

Contents

Supplementary Methods	4
Molecules and Solubility	4
Crystal structure and gas-phase calculations	4
Equation S1. Buckingham potential.....	4
Equation S2. Calculation of the Helmholtz free energy in DMACRYS.....	5
Equation S3. Calculation of entropy from the Helmholtz free energy.....	5
Machine learning regression models:	5
Partial Least Squares Regression (PLSR)	5
Equation S4. Equation for the regression coefficients in PLSR.	6
Random Forest Regression (RF).....	6
Support Vector Regression (SVR).....	7
Figure S1. (A) PLSR; (B) SVR with a soft margin loss function.	8
Machine Learning Model Parameters	9
Statistical Test Formulas.....	9
Equation S5. <i>RMSE</i> and R^2 equations.....	9
Statistical Significance test	9
Equation S6. P value	10
Solubility Challenge Dataset.....	11
Crystal Structure, Molecule Name and Molecular Structure	12
Table S1: Molecular structures, CSD refcodes and chemical names of the 100 molecules	12
Conversion of Experimental and Calculated Values to Log S.....	47
Equation S7. Log S (units referred to mol/L)	47
Equation S8. Theoretical definition.	47
25 Molecule Dataset	47
Chart S1: 25 molecule dataset predictions SMD(HF).....	47
Table S2: Names, CSD refcodes and SMILES for the 25 molecules in dataset DLS-25.	48
Chart S2: 75 molecule dataset predictions SMD(HF).....	49
Chart S3: 100 molecule dataset predictions SMD(HF).....	49
Chart S4: 25 molecule dataset predictions DFT SMD(M06-2X).	50
Chart S5: 75 molecule dataset predictions DFT SMD(M06-2X).	50
Chart S6: 100 molecule dataset prediction DFT SMD(M06-2X).	51
Supplementary Results.....	51

R^2 results	51
Table S3. Cheminformatics descriptors: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation.	51
Table S4. Hartree-Fock energy terms: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation obtained when Hartree-Fock energy terms are used as features in machine learning.	51
Table S6. Hartree-Fock energy terms and Cheminformatics descriptors: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation.	52
Table S5. M06-2X: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation obtained when M06-2X energy terms are used as features in machine learning.	52
Table S7. M06-2X and Cheminformatics descriptors: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation.	52
Table S9. Solubility Challenge dataset: R^2 for the calculated against experimental log S values for ten repetitions of 10-fold cross-validation using cheminformatics descriptors.	52
Table S8. Solubility Challenge dataset: R^2 for the calculated against experimental log S values for the original Solubility Challenge training:test split using cheminformatics descriptors.	52
$RMSE$ results	53
Table S10. Cheminformatics descriptors: average over ten repetitions of the 10-fold cross-validation of $RMSE \pm$ Standard Deviation for the predicted and experimental log S values.	53
Table S11. Hartree-Fock energy terms: average over ten repetitions of the 10-fold cross-validation of $RMSE \pm$ Standard Deviation for the predicted and experimental log S values obtained when HF energy terms are used as features in a machine learning model.	53
Table S13. Hartree-Fock energy terms and Cheminformatics descriptors: average $RMSE \pm$ Standard Deviation over ten repetitions of the 10-fold cross-validation for the predicted and experimental log S values.	53
Table S12. M06-2X energy terms and Cheminformatics descriptors: average over ten repetitions of the 10-fold cross-validation of $RMSE \pm$ Standard Deviation for the predicted and experimental log S values.	53
Table S14. M06-2X energy terms: average over ten repetitions of the 10-fold cross-validation of $RMSE \pm$ Standard Deviation for the predicted and experimental log S values obtained when M06-2X energy terms are used as features in a machine learning model.	54
Table S15. Solubility Challenge dataset: $RMSE$ for the calculated against experimental log S values for ten repetitions of 10-fold cross-validation using cheminformatics descriptors.	54
Table S16. Solubility Challenge dataset: $RMSE$	54
Statistical Significance Test	55
BOX S1: P-value (statistical significance at $P = 0.05$) of the performance of the $RMSE$ scores for the different regression models for the scaled dataset by using mean/stdev.	55

BOX S2: P-value (statistical significance at $P = 0.05$) of the performance of the <i>RMSE</i> scores for the different regression models for the scaled dataset by Principal Components.	56
BOX S3: P-value (statistical significance at $P = 0.05$) of the performance of the <i>RMSE</i> scores for the different regression models for the row dataset.	57
Variable Importance.....	58
Table S17: Top 10 results of variable importance for different descriptors and dataset.....	58
References.....	58

Supplementary Methods

Molecules and Solubility

Even for identical chemical names, the SMILES strings found in various well-regarded databases may imply subtly different chemical structures. Typically, variants may differ in stereochemistry, protonation state and in the treatment of aromaticity which is sometimes expressed as alternating single and double bonds, rather than as canonically aromatic structures. Such variations affect the descriptors calculated by CDK.

Crystal structure and gas-phase calculations

The Buckingham potential is:

$$U_{rep-disp}^{MN} = \sum_{i \in M, k \in N} A_{ik} \exp(-B_{ik} R_{ik}) - \frac{C_{ik}}{R_{ik}^6}$$

Equation S1. Buckingham potential.

where i and k are atoms in molecules M and N , with the fitted values A_{ik} , B_{ik} and C_{ik} being characteristic of the interaction between the relevant atom types and R_{ik} being the distance separating atoms i and k . A_{ik} , B_{ik} and C_{ik} are fitted to experimental results.

Geometry optimizations were carried out in duplicate using M06-2X/6-31G* and HF/6-31G*, starting from hydrogen-normalized versions of the crystal structure monomer geometries.¹ All calculations were done using G09's "ultrafine" integral grid (containing 99 radial shells and 590 angular points per shell) because it is known that the M06-2X functional is sensitive to integral grid spacing.²

The Helmholtz free energy free energy is calculated as follows:

$$F = U + \frac{1}{2} \sum_i h\nu_i + kT \sum_i \ln(1 - e^{-\frac{h\nu_i}{kT}})$$

Equation S2. Calculation of the Helmholtz free energy in DMACRYS.

where F is the Helmholtz free energy, U is the energy of the stationary lattice, ν_i are the frequencies of the normal modes, k is the Boltzmann constant and T is the absolute temperature.³

$$\left(\frac{\delta F}{\delta T}\right)_V = -S$$

Equation S3. Calculation of entropy from the Helmholtz free energy.

The partial derivative of the Helmholtz free energy with respect to temperature at constant volume gives the negative of the entropy.

Machine learning regression models:

Partial Least Squares Regression (PLSR)

In the given dataset of n observations (druglike molecules), the dataset is $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where x_i ($i = 1, \dots, n$) is a vector of descriptors and y_i is the property or activity of interest, here log S [Figure 1 A]. The given data D are split into training and test sets, where the training set of X is used in order to fit to the PLSR model. The predictions for new observations are based on the training set by decomposing the data into singular vectors. For this, first, the data matrices X and Y are decomposed using singular value decomposition of their cross product matrix S , where $S = X^T Y$. The singular value decomposition of S is $SVD(S) = WAC^T$, which is the main analytical tool in PLSR. In PLSR, this kind of decomposition is also known as eigenvalue decomposition.⁴ The left (*i.e.* W) and right (*i.e.* C) singular vectors are used as weight matrices W and C of X and Y , respectively, to obtain

scores T ($T = XW$) and U ($U = YC$) that explain the data [Figure S1 A]. It is not necessary to calculate the score matrix of Y in regression analysis, but it is still often used for interpretation. Next, loadings (*i.e.* P) are calculated by regressing against the same vector T , $P = X^T T$. These matrices will be normalised by subtracting the loadings from the original data matrix. The complete steps are iterative in order to retrieve the estimate of the components. Afterwards the scores T are used to calculate the matrix of regression coefficients B (as in Equation S4), which is converted back to the realm of the original variables by pre-multiplying by R ; $R = [W(P^T W)^{(-1)}]$.

$$B = R (T^T T)^{(-1)} T^T Y$$

Equation S4. Equation for the regression coefficients in PLSR.

Random Forest Regression (RF)

In the given dataset D we have n instances, here molecules, used for the tree-building process that constitutes the training set. The random forest is an ensemble of decision trees $\{T_1(X), \dots, T_b(X)\}$, each tree generated by stochastic recursive partitioning of a bootstrap sample of the training set. As the molecules progress through the tree, they are partitioned into increasingly homogeneous groups, so that each terminal node of the decision tree is associated with a group of molecules with similar solubility. Each split within a tree is created based on the best partitioning of the bootstrap sample, according to the Gini criterion, that is possible using any of a randomly chosen subset of $mtry$ descriptors. This random subset is freshly chosen for each node. If $ntree$, the number of trees in the forest, is held constant then $mtry$ is the only parameter that needs to be optimised. For each tree, approximately one third of the training set molecules do not appear in that tree's bootstrap sample, and constitute the so called out-of-bag data; conversely, every molecule is out-of-bag for about a third of the trees.

In the prediction phase the test molecules are passed through the trees built from the training data. Each tree provides the output $Y_1^{pred} = T_1(X)$,..., $Y_b^{pred} = T_b(X)$, where Y_b^{pred} contains the

prediction for the test molecules by the b^{th} tree. Lastly, the outputs of each tree for each given molecule are averaged to produce the random forest's final prediction of log S for that compound.

Different kinds of experimental design are possible. In one possible design, only out-of-bag predictions are carried out and each molecule is predicted only by those trees for which it was not part of the bootstrap training sample. In another design, where the test set is entirely external, the trees are constructed from the training data and every tree will be used to predict every test compound. In the 10-fold cross-validation design used in this work, a random forest is constructed from nine of the 10-folds and used to predict the solubilities for the molecules constituting the tenth fold; this process is then repeated cyclically with each fold successively being predicted by a random forest constructed from the other nine.

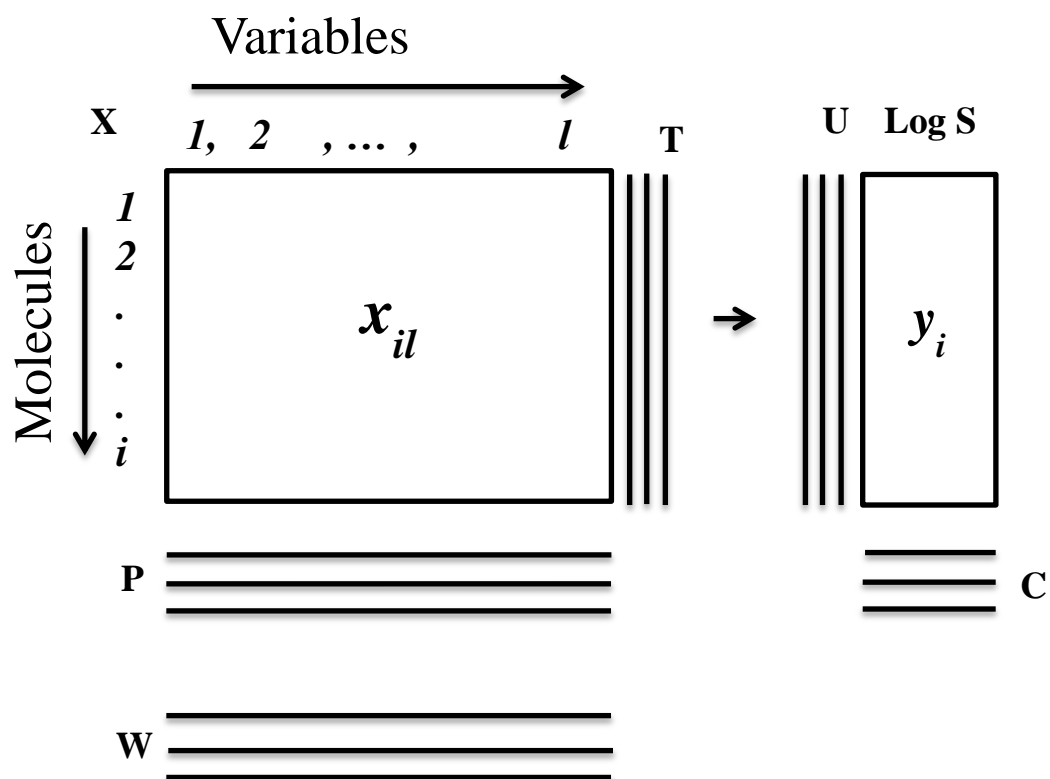
Support Vector Regression (SVR)

To compute the linear regression of the given training data X , SVR approximates the function in the following way: $f(x_i) = \omega^t x_i + B$, where ω is a vector of weights and B is the constant coefficient.

In order to estimate the function's deviation from the true one, SVR uses a loss function

$L(Y, f(X, \omega), \varepsilon)$ that was introduced by Vapnik [Figure S1 B]. SVR uses an ε -insensitive loss function in order to capture the deviation of $f(X, \omega)$ from the actual y_i for the complete training set; this deviation should be at most ε in magnitude. Moreover, the SVR algorithm tries to reduce model complexity by minimizing the weights $||\omega||^2$. This is a very stringent rule; it implies that a function f exists that approximates (X, Y) with precision ε . This is not always the case, where the situation is not so stringent, slack variables, ξ_i, ξ_i^* ; $i = 1, \dots, n$, are introduced to provide flexibility to the model for each of the n molecules.⁵

A: Partial Least Square Regression



B: Support Vector Regression

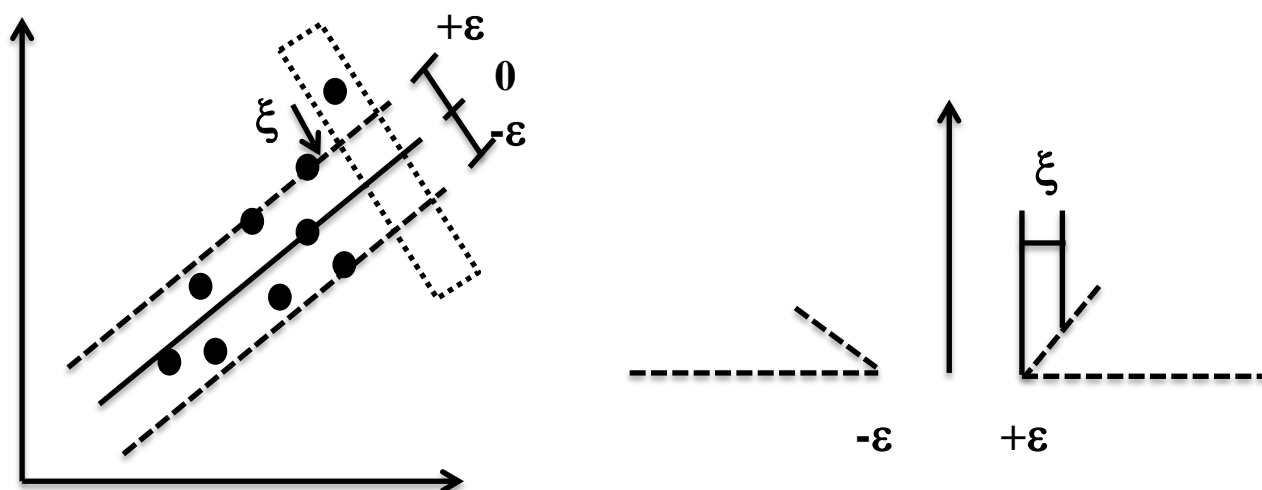


Figure S1. (A) PLSR; (B) SVR with a soft margin loss function.

Machine Learning Model Parameters

1. For PLSR the only parameter that was optimised is the number of components '*ncomp*' which ranged from 1-20.
2. In RF if *ntree*, the number of trees in the forest, is held constant then *mtry* is the only parameter that needs to be optimised. The range of parameters are: '*mtry*' (2-123) and '*ntree*' (set at: 500)
3. In SVR we used the radial basis kernel function where two parameters play important roles: the capacity parameter *C* (for which we tried twenty different values varying between 0.25 and 131072), and *sigma* (set at: 0.0112).

For parameter optimisation we performed 10-fold cross-validation within the training set. The parameter optimisation was done using the CARET package.

Statistical Test Formulas

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{obs} - y^{pred})^2}$$

$$R^2 = \left(\frac{\sum (x_{pred} - \bar{x})(y_{pred} - \bar{y})}{\sqrt{\sum (x_{pred} - \bar{x})^2 \sum (y_{pred} - \bar{y})^2}} \right)^2$$

Equation S5. *RMSE* and R^2 equations (R^2 here is the square of the Pearson's correlation coefficient not the coefficient of determination). Where *n* is the number of molecules, y^{obs} is the observed output and y^{pred} is the predicted output, \bar{x} is the mean value of *x*, \bar{y} is the mean value of *y*.

Statistical Significance test

The permutation test is widely used technique in various research areas such as in bioinformatics and chemoinformatics where the question is how well algorithm A performed compared with algorithm B on a particular problem characterised by a data set *D*.⁶ By using the permutation test one

can calculate exact P-values for the commonly used 10-fold cross-validation methods by using fewer assumptions about the distribution of a paired difference. In this study we are using a permutation test,⁷ to test for significantly different performance (*via RMSE*) between the two regression models by their P-values.

$$P = \frac{n}{N}$$

Equation S6. P value

where n is the number of permutations of the mean difference in the performance of two regression models that can be more extreme than the observed mean difference and N is the total number of possible reassignments of the paired differences given the results. In more detail, the procedure consists of the following steps:

1. A given paired-difference (B_0) of *RMSE* scores obtained by different regression models is given by $B_0 = (R_A^1 - R_B^1) + (R_A^2 - R_B^2) + \dots + (R_A^{10} - R_B^{10})$ where R_A^1 is the *RMSE* scores for test set predictions made by model A for each fold (1 ... 10) in the 10-fold cross-validation.
2. For this statistical test, 1024 permutations are created *via* all 2^{10} combinations: $B_p = \pm(R_A^1 - R_B^1) \pm (R_A^2 - R_B^2) \pm \dots \pm (R_A^{10} - R_B^{10})$.
3. The rank of true difference in the performance B_0 is used as an indicator of the p-values among the 1024 permutations. The P-value is computed as: $P = \frac{n}{1024}$ where n is the number of permutations which have $|B_p| \geq |B_0|$.

Variable Importance

The variable importance can be calculated with a model-dependent or model-independent method. A model-dependent method has the advantage of using information from the model performance, for example in algorithms such as Random Forest. Here, we use the CARET package to evaluate the variable importance “*varImp*” for Random Forest.^{8 9} Irrespective of the method of calculation, the variable importance scores are scaled to a maximum of 100. The variable importance is calculated as the average difference between a conventional out-of-bag prediction and a second “noised up”

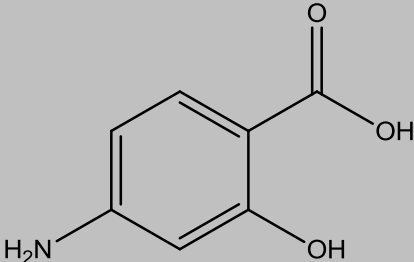
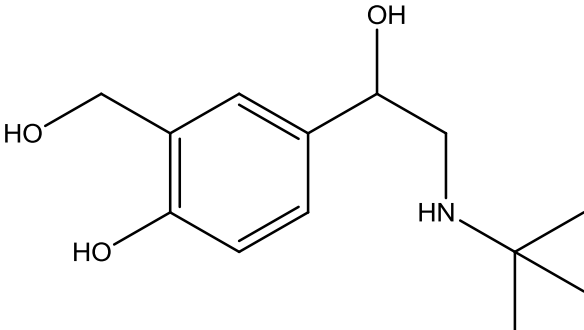
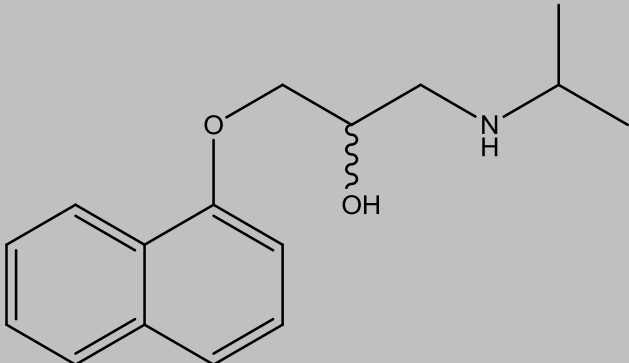
prediction in which a single descriptor has its values permuted between molecules. The most important descriptor is then the one giving the largest reduction in accuracy when noised up. The variable importance for the descriptors used in this study are in Table S14 and definitions of the CDK descriptor names can be found in¹⁰.

Solubility Challenge Dataset

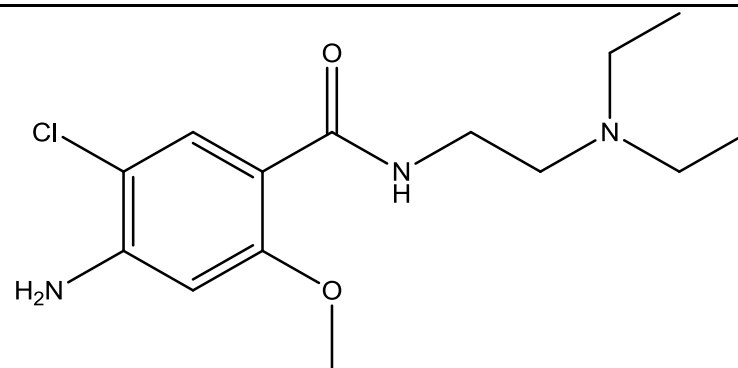
As a benchmark, we also used our descriptor-based methodology retrospectively to replicate the Solubility Challenge itself. We used the Solubility Challenge dataset as a benchmark in this work to directly compare our method to others and to judge the relative difficulty of our 100-molecule dataset against that of the standard Solubility Challenge set. The Solubility Challenge dataset comes from work by Llinas *et al.*,¹¹ where solubilities of 122 compounds are reported from the CheqSol method. The molecules were selected on the basis of commercial availability and must contain an ionisable group. Hence we trained our models with the 94 training set solubilities from the original Solubility Challenge,¹¹⁻¹² and tested on the 28 molecules of the test set (more specifically, there are 94 useable quantified solubilities amongst the 100 molecules of the original training set, and 28 amongst the 32 compounds in the canonical test set). ChemSpider SMILES were used for 90 of the training set molecules and for all 28 test set compounds; for 5-bromogramine, cimetidine, pindolol and phenobarbital we instead took the SMILES from the Solubility Challenge web site^{12a} in order to obtain the desired neutral protonation state. Since 60 molecules of the training set and 24 of the test set had no suitable crystal structure in the CSD, we could not calculate energy descriptors for the Solubility Challenge set.

Crystal Structure, Molecule Name and Molecular Structure

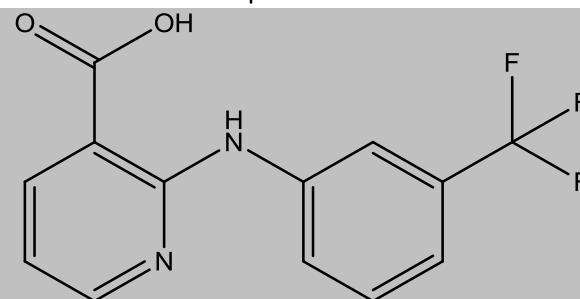
Table S1: Molecular structures, CSD refcodes and chemical names of the 100 molecules used in this study. The SMILES for this DLS100 dataset can be found in the zip file of solubility datasets and scripts that forms part of the Supporting Information.

Number	Crystal structure	Molecule name	Molecule structure
1	TAYGAC	Nadolol	
2	BHHPHE	Salbutamol	
3	IMITON	Propranolol	

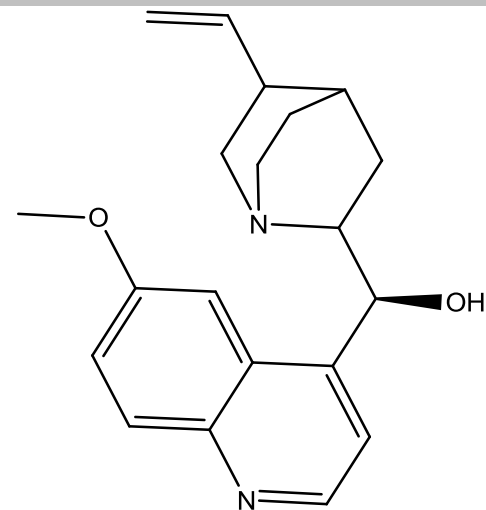
4 METPRA Metoclopramide



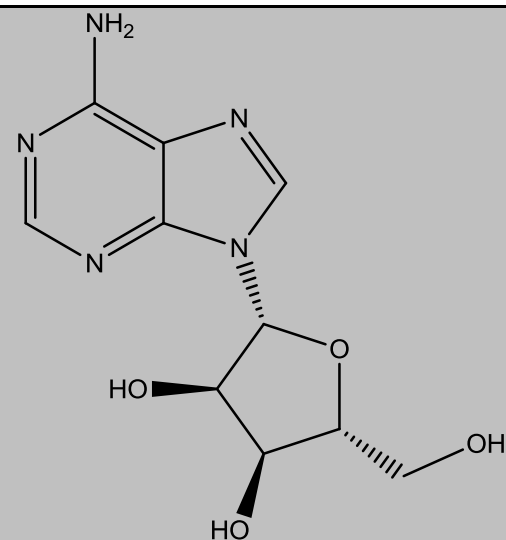
5 NIFLUM10 Niflumic acid



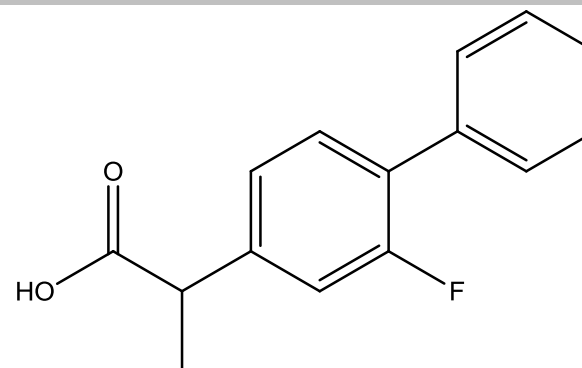
6 BOMDUC Quinidine



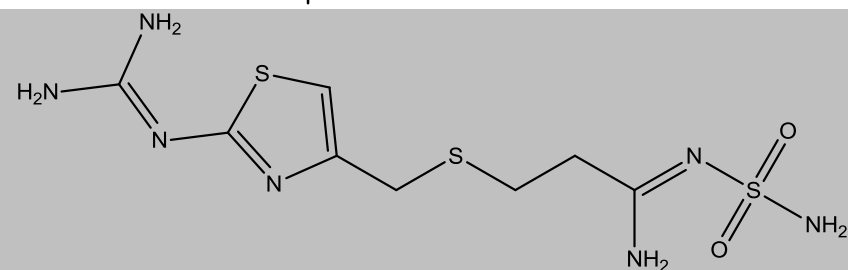
7 ADENOS10 Adenosine



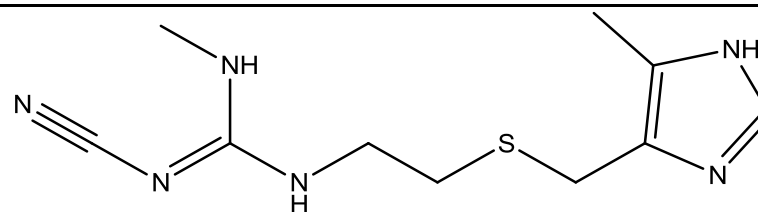
8 FLUBIP Flurbiprofen



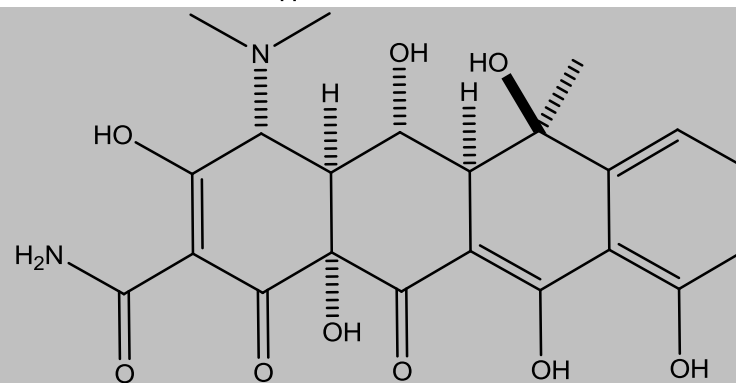
9 FOGVIG02 Famotidine



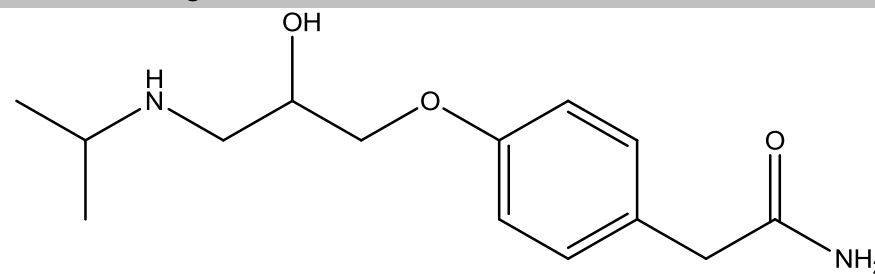
10 CIMETD Cimetidine



11 OXYTET Oxytetracycline



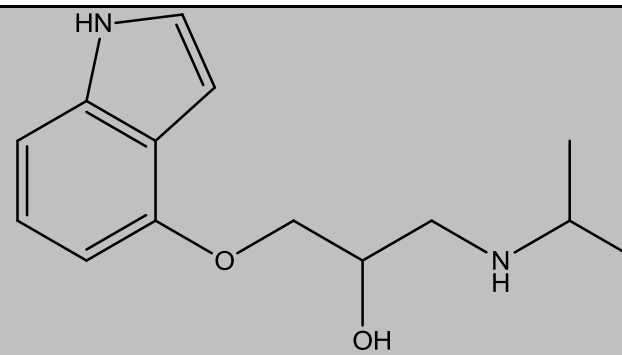
12 CEZVIN (RS)-Atenolol



13

PINDOL

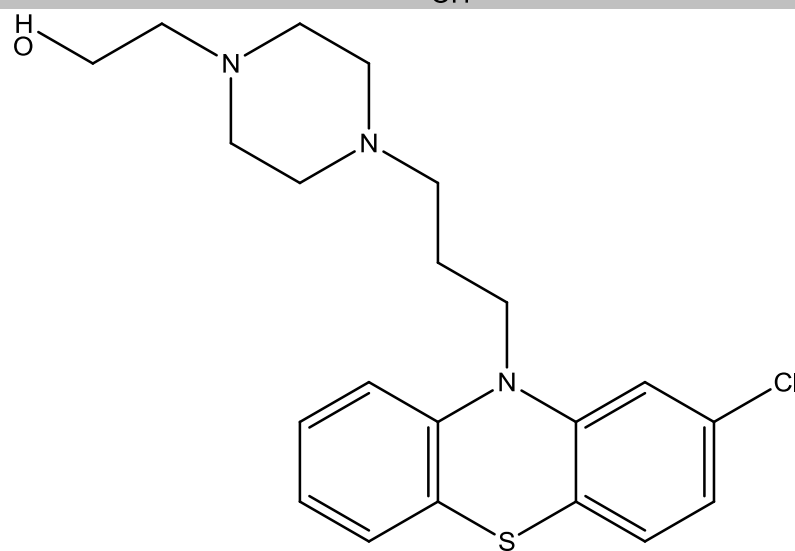
Pindolol



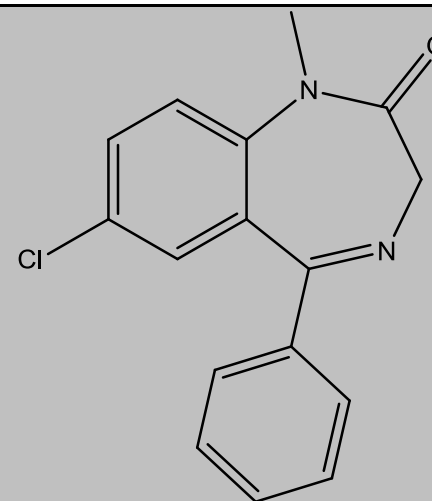
14

PERPAZ

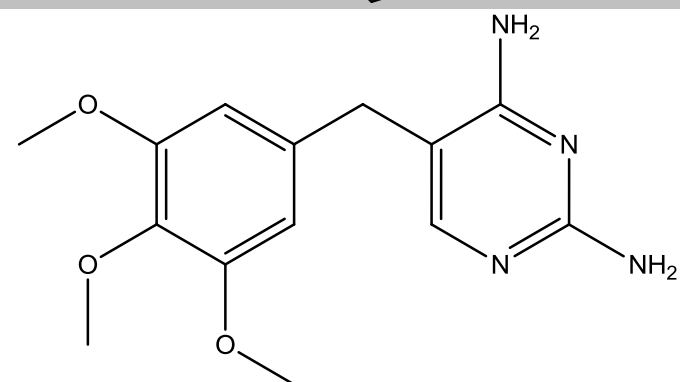
Perphenazine



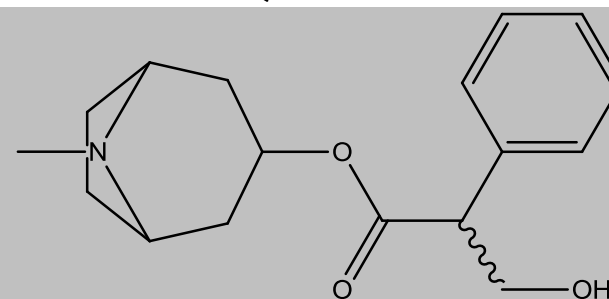
15 DIZPAM10 Diazepam



16 AMXBPM10 Trimethoprim



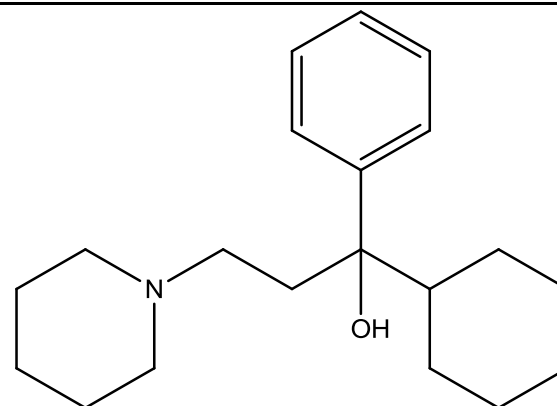
17 WALPIJ Atropine



18

THEXPL

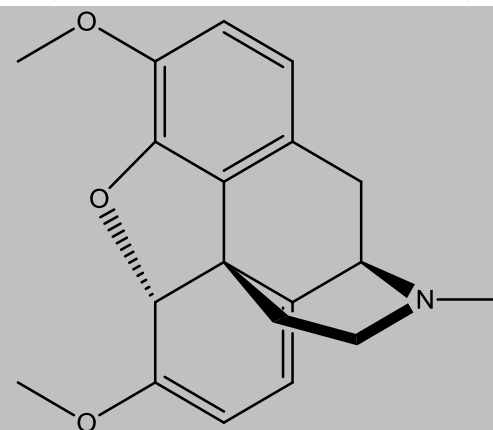
Trihexyphenidyl



19

TICTUU

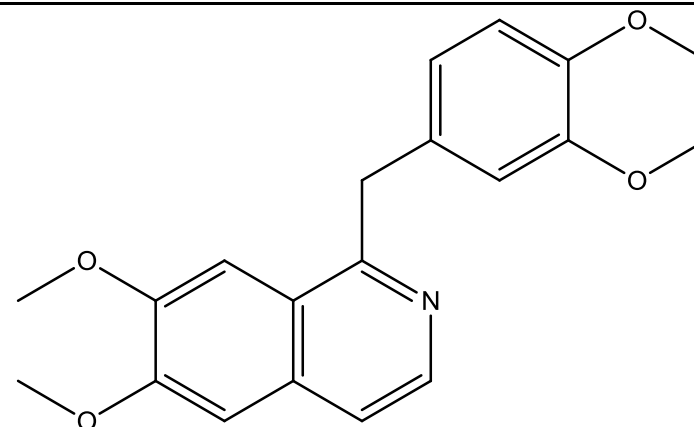
Thebaine



20

MVERIQ01

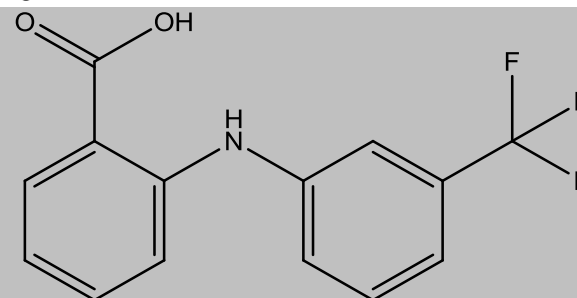
Papaverine



21

FPAMCA

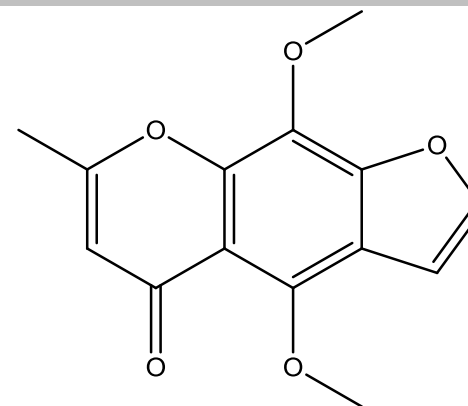
Flufenamic acid



22

KHELIN

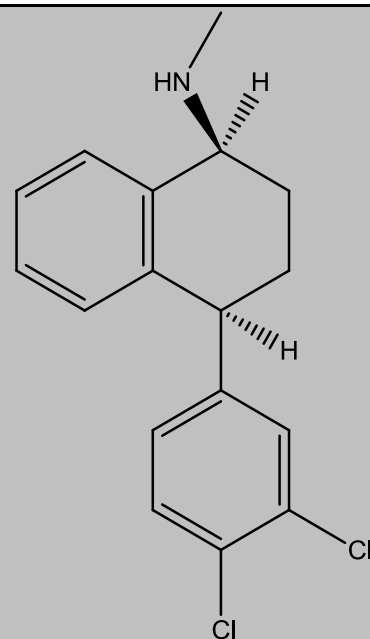
Khellin



23

CUTPEN

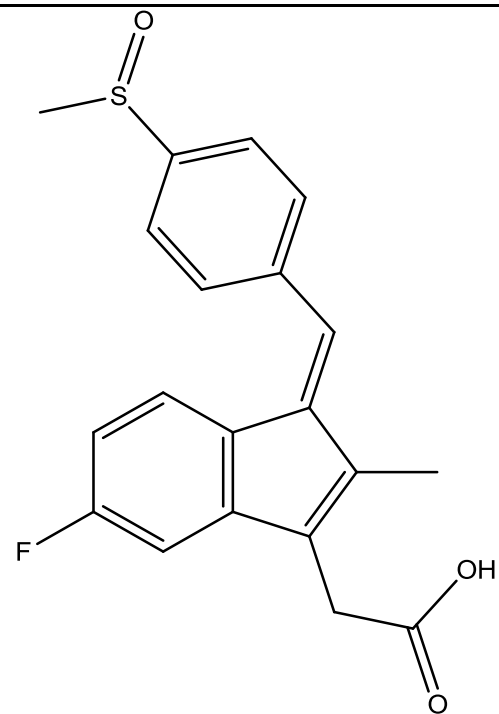
Sertraline



24

DOHREX

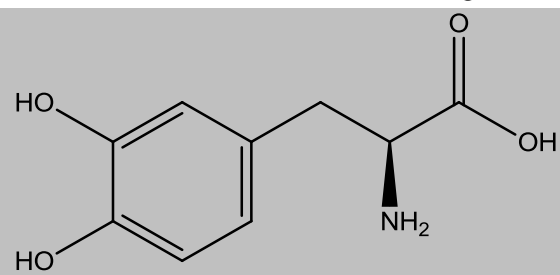
Sulindac



25

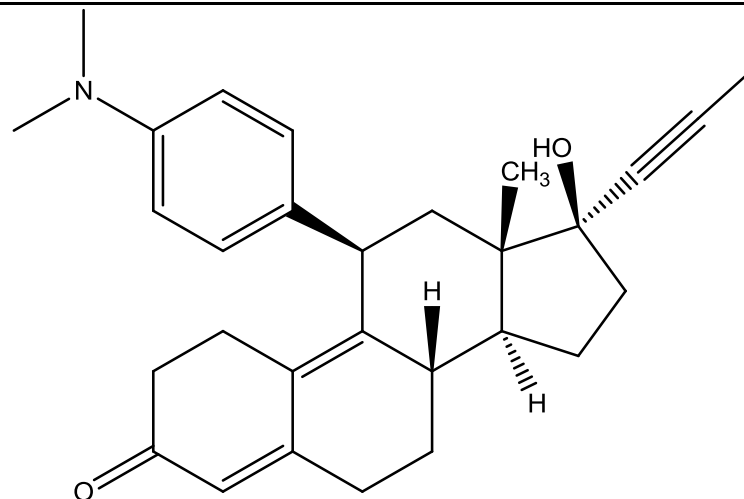
LDOPAS03

L-DOPA (Levodopa)



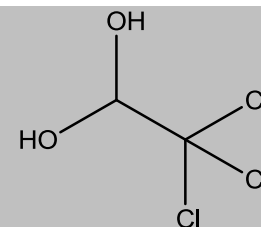
ZIDLED

Mifepristone



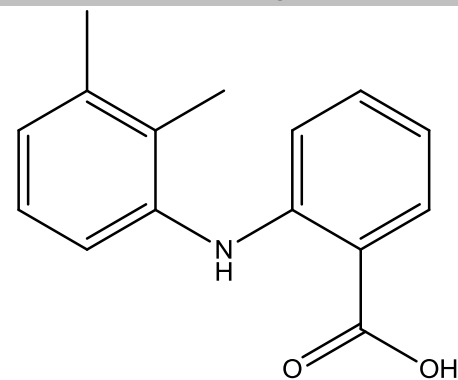
CHORLH01

Chloral Hydrate

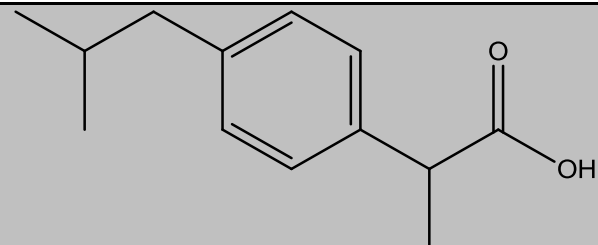


XYANAC

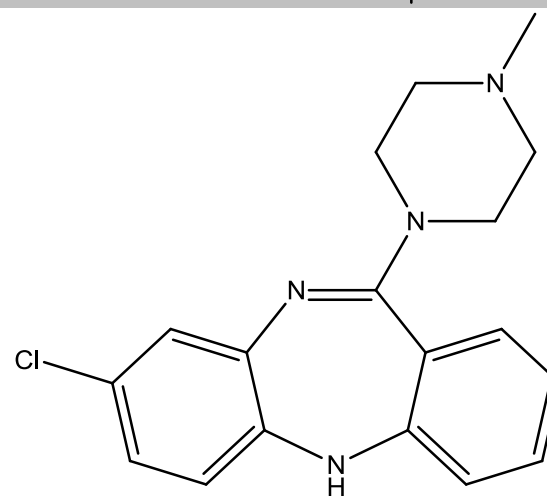
Mefenamic acid



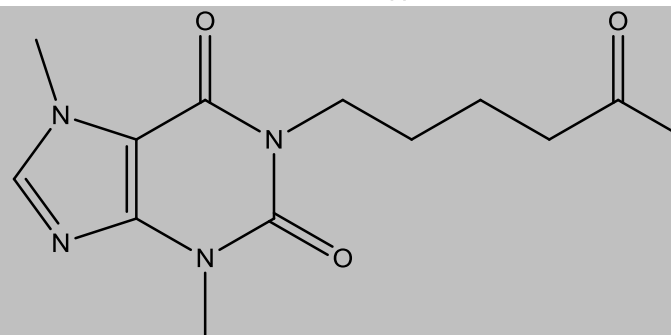
29 IBPRAC01 Ibuprofen



30 NDNHCL01 Clozapine



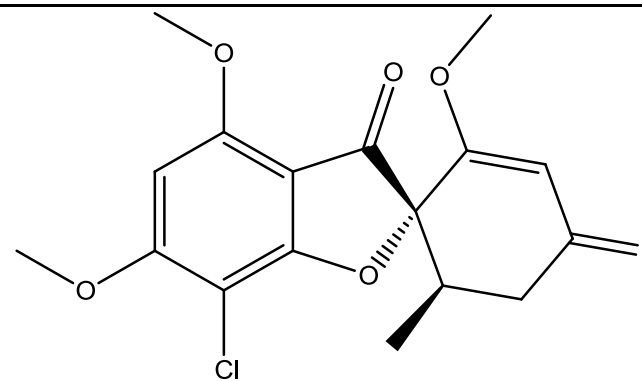
31 JAKGEH Pentoxifylline



32

GRISFL

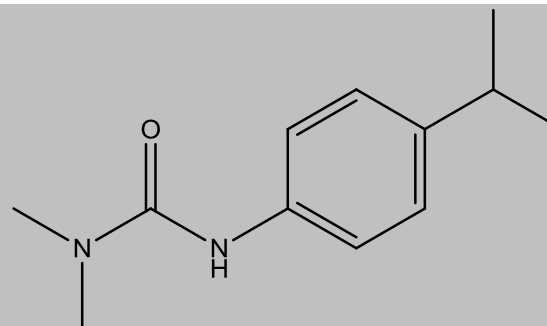
Griseofulvin



33

JODTUR01

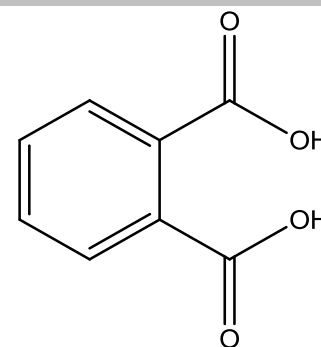
Isoproturon



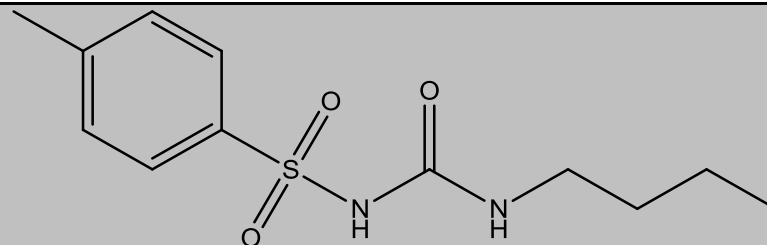
34

PHTHAC01

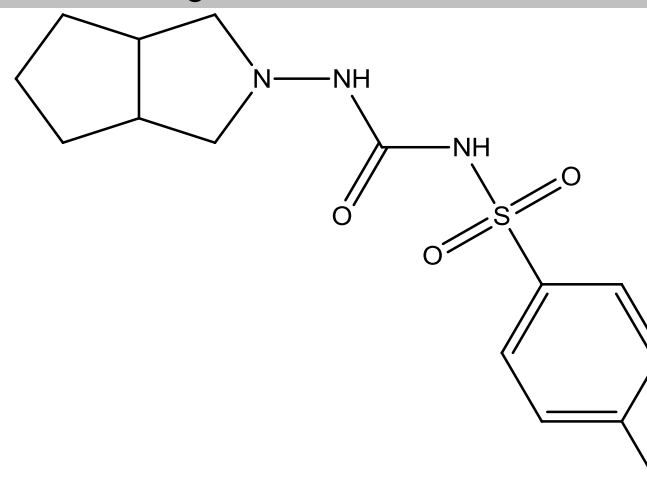
Phthalic acid



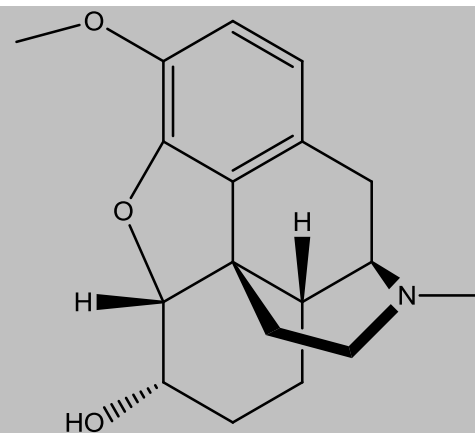
35 ZZZPUS02 Tolbutamide



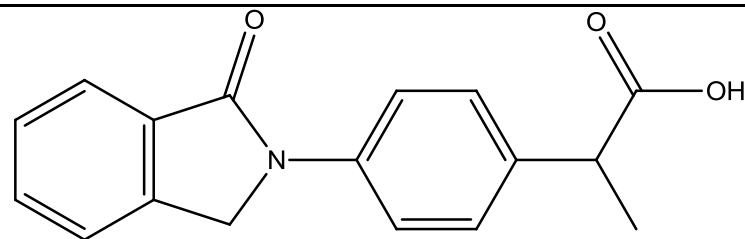
36 SUVGUL Gliclazide



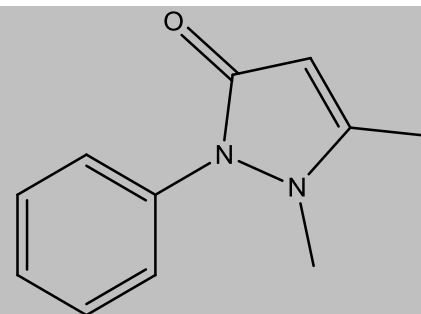
37 ZZZTSE03 Codeine



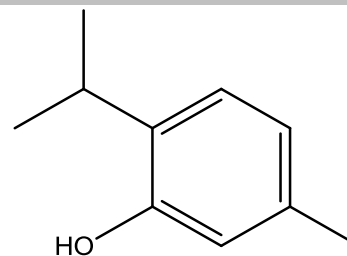
38 LEKMET Indoprofen



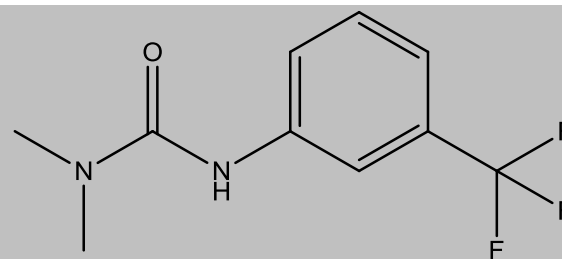
39 ANTPYR10 Antipyrine



40 IPMEPL Thymol



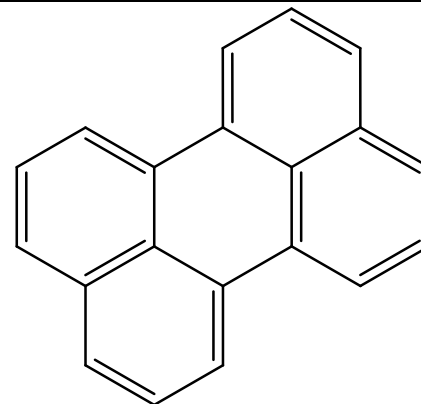
41 HODHIS Fluometuron



42

PERLEN05

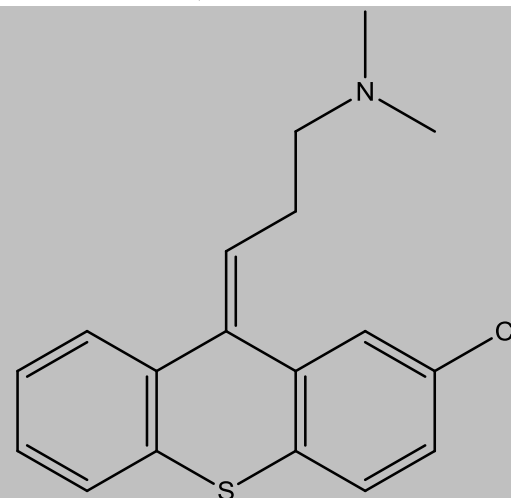
Perylene



43

CMAPTXX

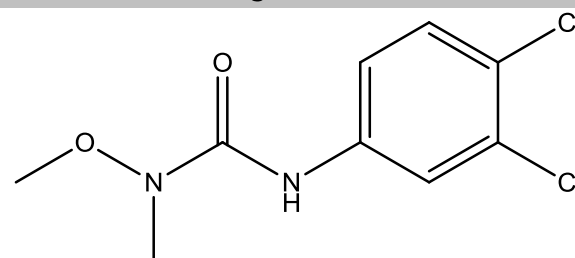
Chlorprothixene



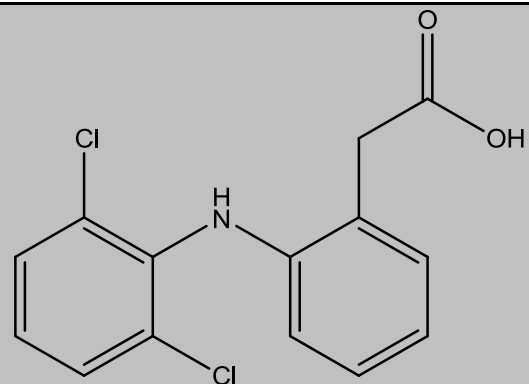
44

WAMXUD

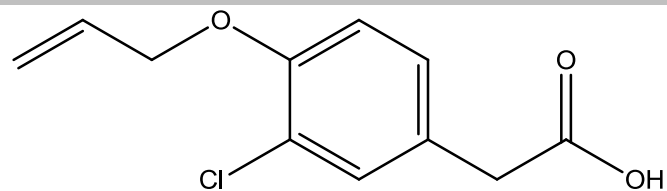
Linuron



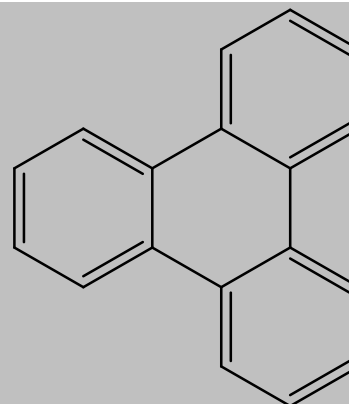
45 SIKLIH01 Diclofenac



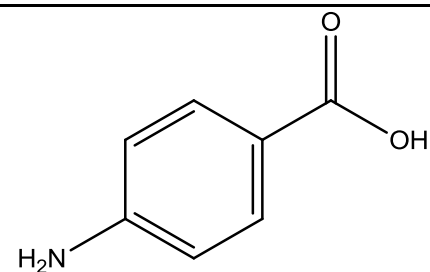
46 FICJAC Alclofenac



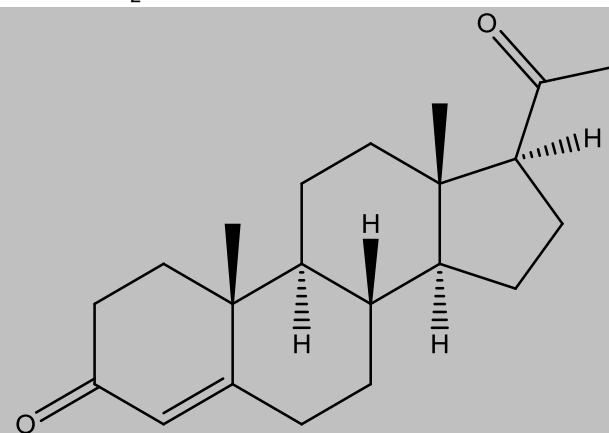
47 TRIPHE11 Triphenylene



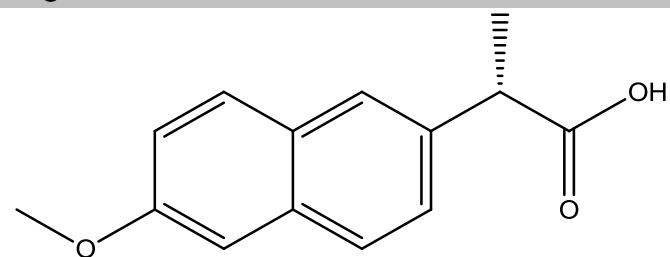
48 AMBNAC04 4-Aminobenzoic acid



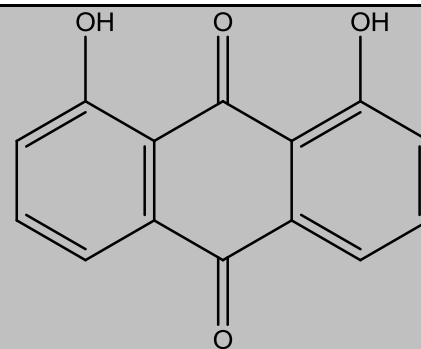
49 PROGST12 Progesterone



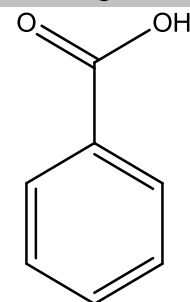
50 COYRUD11 Naproxen



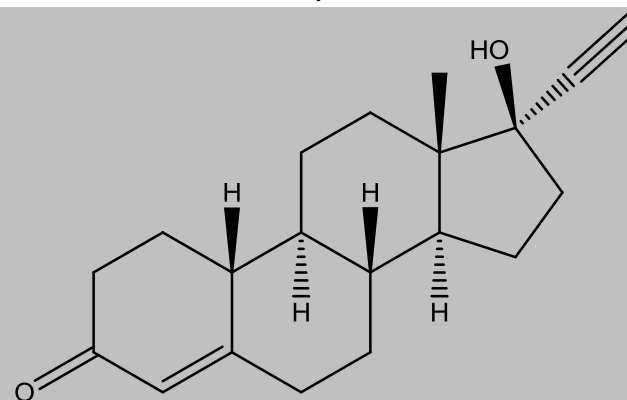
51 DHANQU06 1,8-Dihydroxyanthraquinone



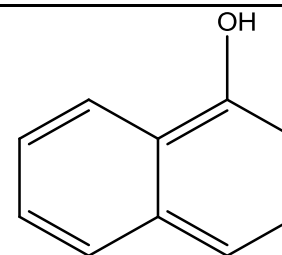
52 BENZAC02 Benzoic acid



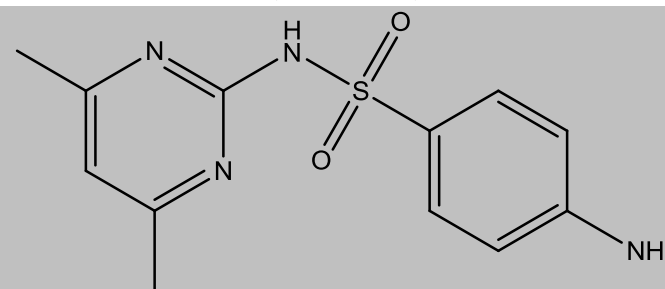
53 NETIND01 Norethisterone



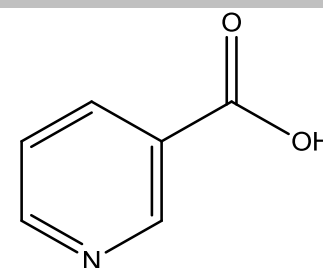
54 NAPHOL01 1-Naphthol



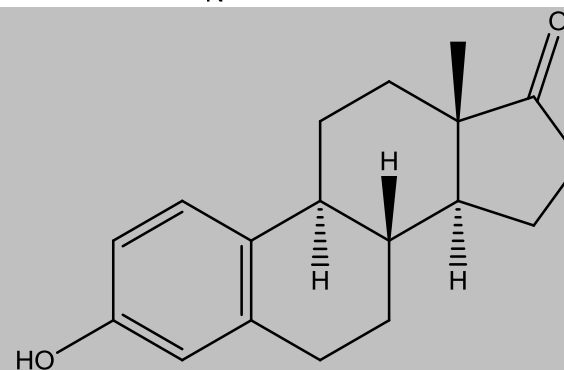
55 SLFNMD01 Sulfamethazine



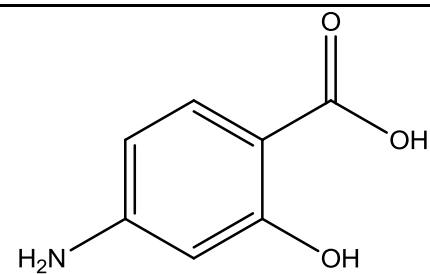
56 NICOAC02 Nicotinic acid



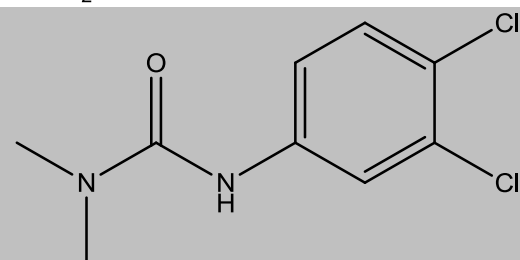
57 ESTRON14 Estrone



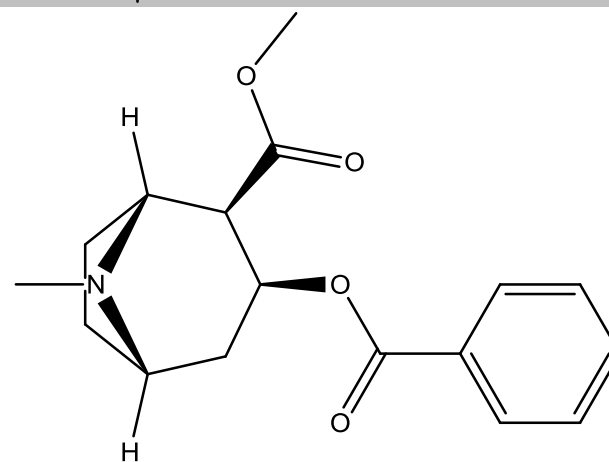
58 AMSALA01 4-Aminosalicylic acid



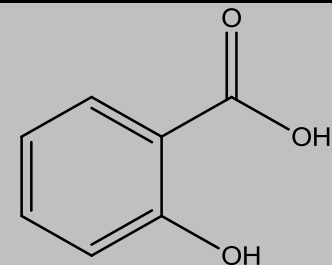
59 CLPHUR02 Diuron



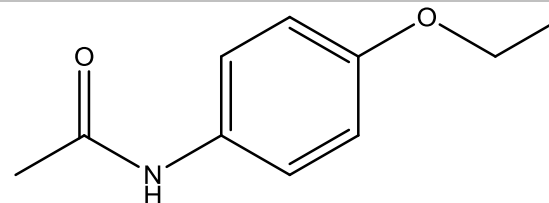
60 COCAIN10 Cocaine



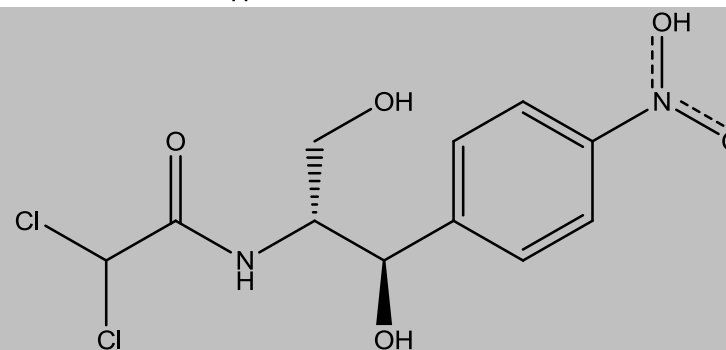
61 SALIAC Salicylic acid



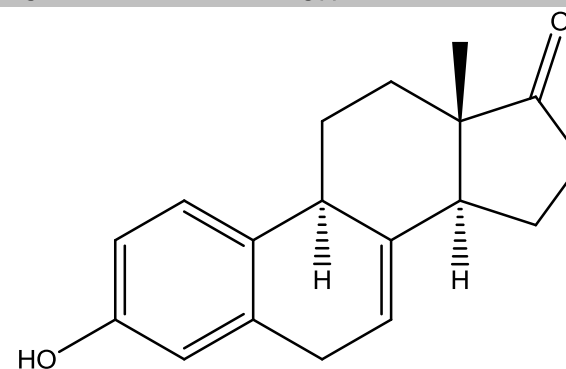
62 PYRAZB21 Phenacetin



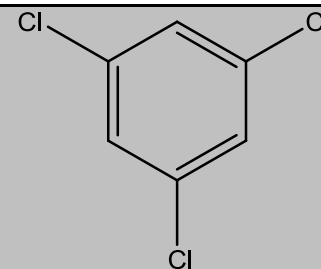
63 CLMPCL02 Chloramphenicol



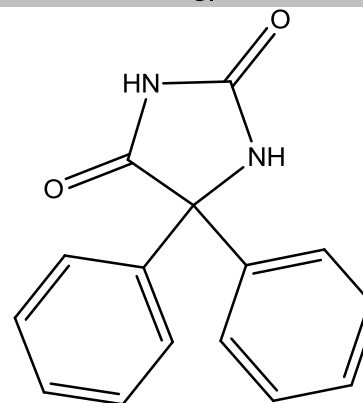
64 GODTIC Equilin



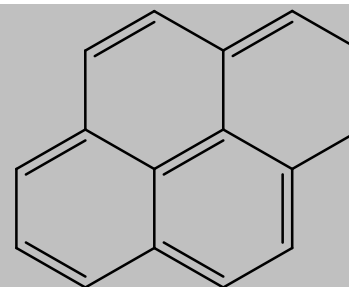
65 TCHLBZ 1,3,5-trichlorobenzene



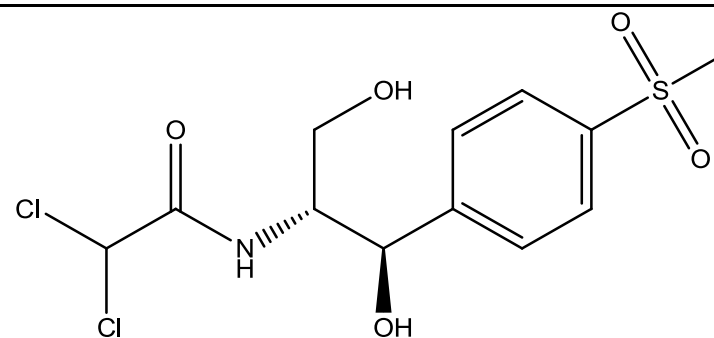
66 PHYDAN01 5,5-Diphenylhydantoin



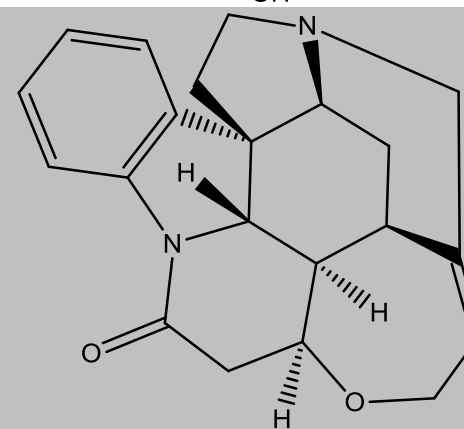
67 PYRENE07 Pyrene



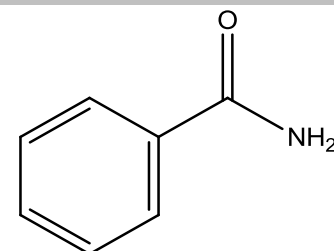
68 CABCIR01 Thiamphenicol



69 ZZZUEE04 Strychnine



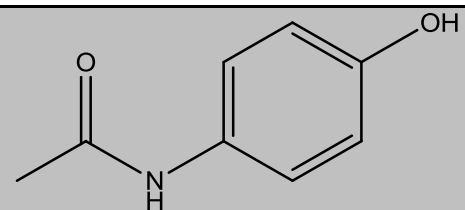
70 BZAMID02 Benzamide



75

HXACAN04

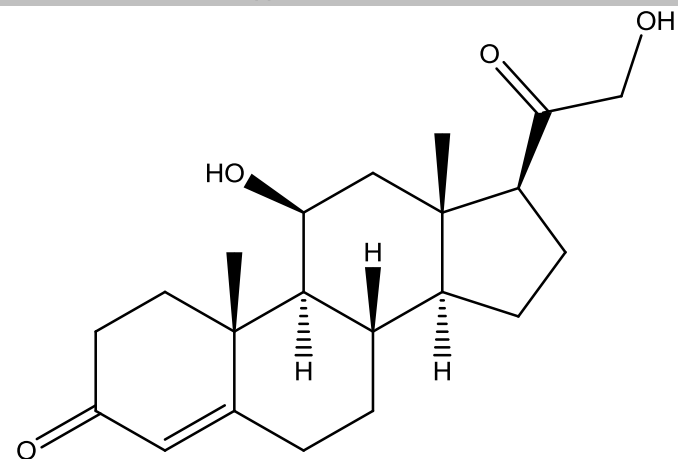
Paracetamol



76

CORTIC

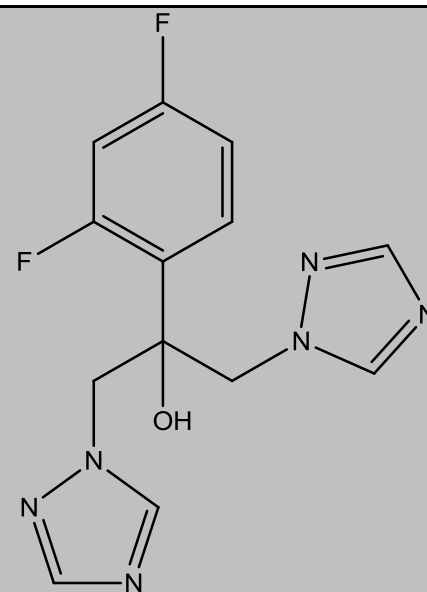
Corticosterone



77

IVUQOF

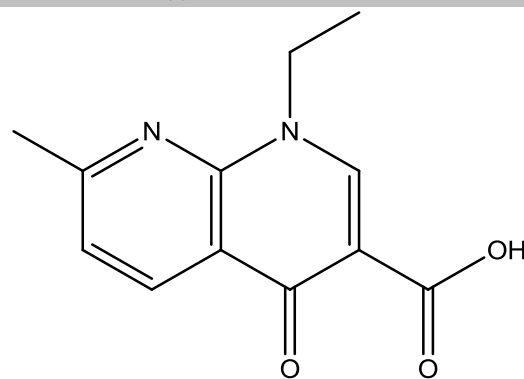
Fluconazole



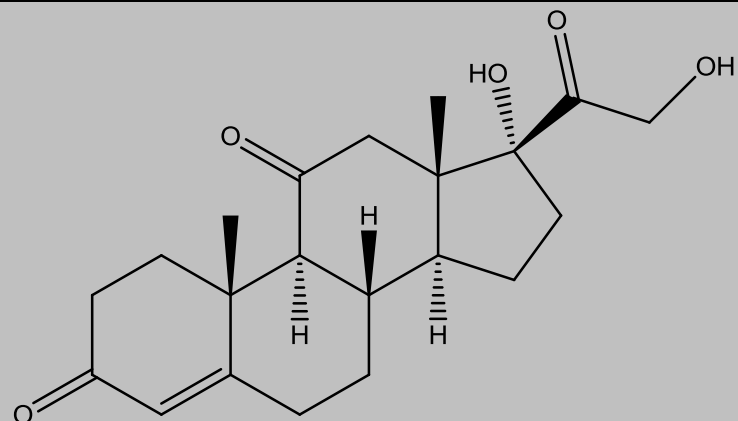
78

NALIDX01

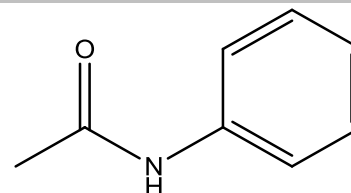
Nalidixic acid



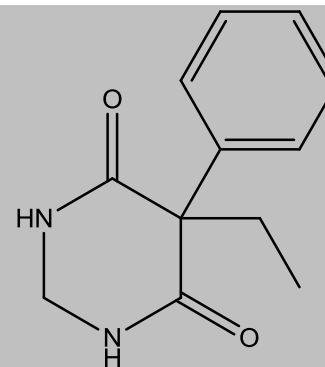
79 DHPRTO02 Cortisone



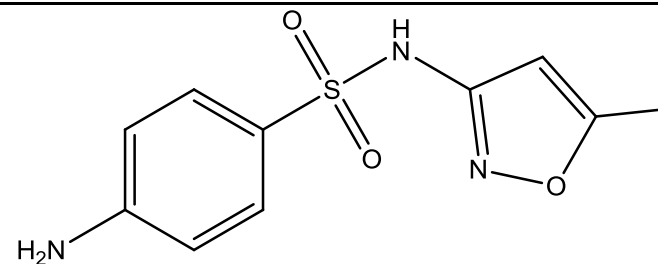
80 ACANIL01 Acetanilide



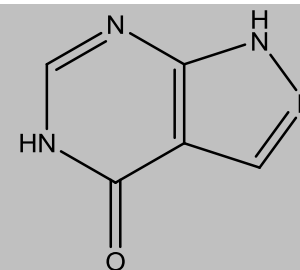
81 EPHPMO Primidone



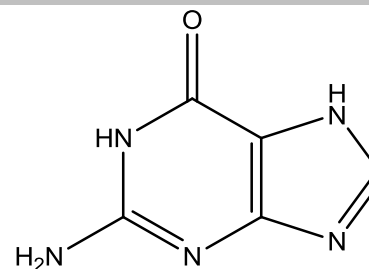
82 SLFNMB01 Sulfamethoxazole



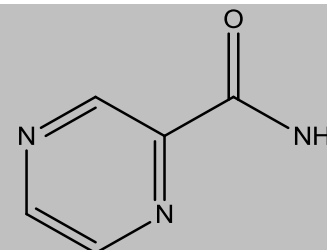
83 ALOPUR Allopurinol



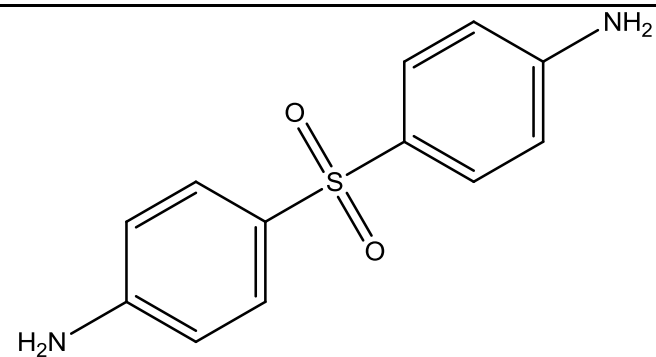
84 KEMDOW Guanine



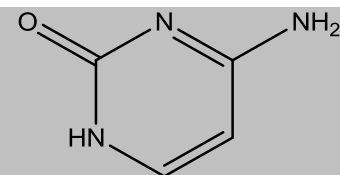
85 PYRZIN Pyrazinamide



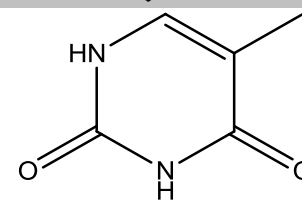
86 DAPSUO03 Dapsone



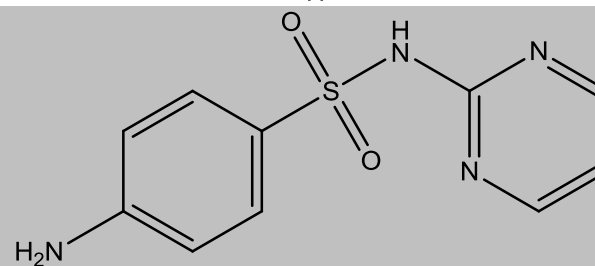
87 CYTSIN01 Cytosine



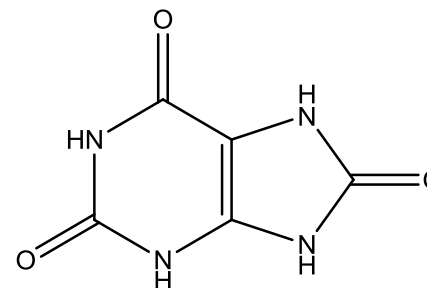
88 THYMIN01 Thymine



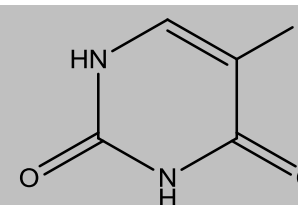
89 SULDAZ01 Sulfadiazine



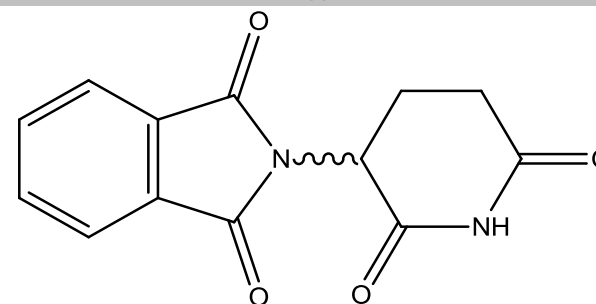
90 URICAC Uric acid



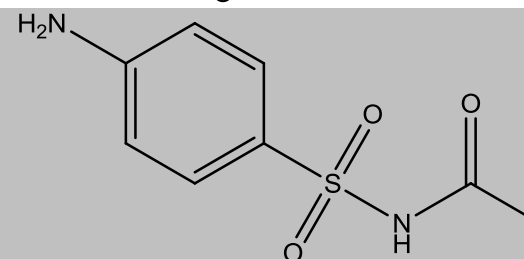
91 FURACL02 5-Fluorouracil



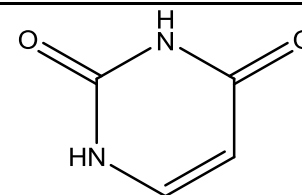
92 THALID03 Thalidomide



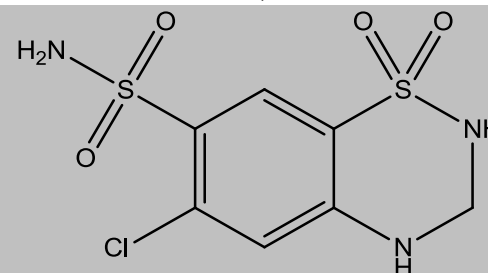
93 SLFNMG01 Sulfacetamide



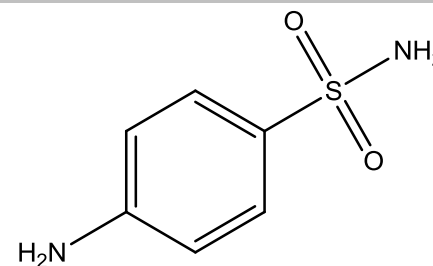
94 URACIL Uracil



95 HCSBTZ04 Hydrochlorothiazide



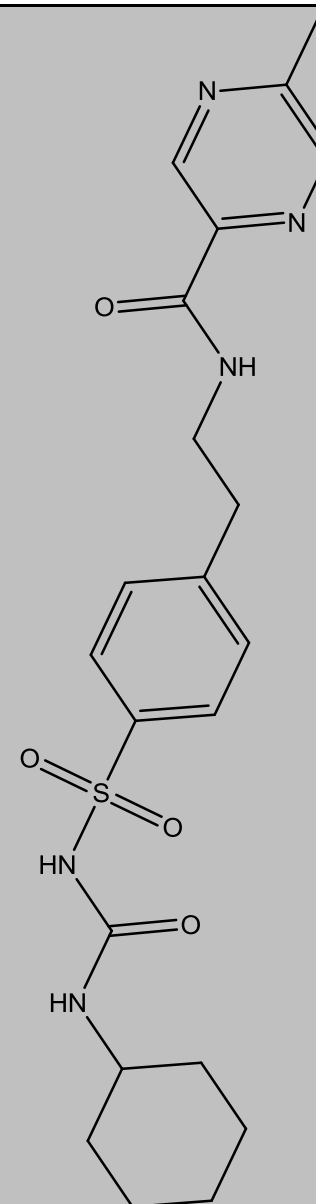
96 SULAMD01 Sulfanilamide



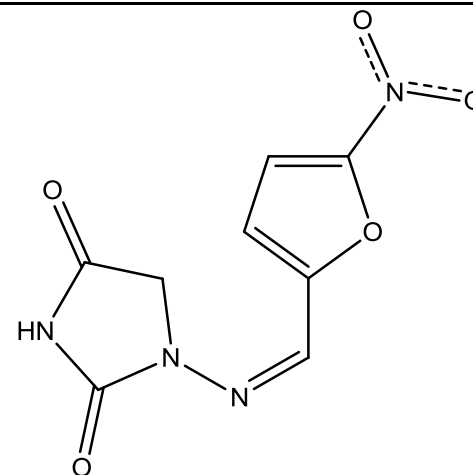
97

SAXFED

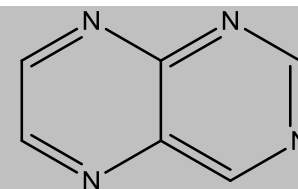
Glipizide



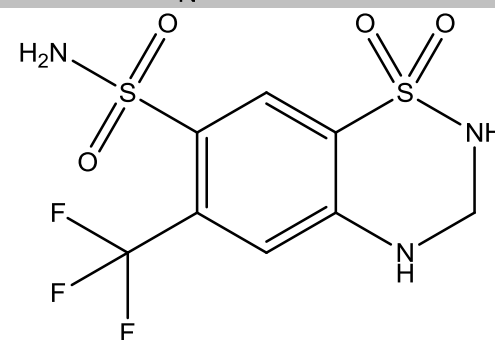
98 LABJON01 Nitrofurantoin



99 PTERID11 Pteridine



100 EWUHAF01 Hydroflumethiazide



Conversion of Experimental and Calculated Values to Log S

For experimental solubility values, log S is found as follows:

$$\text{Log } S = \text{Log}(\text{solubility in mol/L})$$

Equation S7. Log S (units referred to mol/L)

We convert the free energy of solution to log S values:

$$\text{Log } S = \frac{\Delta G_{\text{solution}}}{-2.303RT}$$

Equation S8. Theoretical definition.

where R is the universal gas constant and T is the absolute temperature.

25 Molecule Dataset

25 Molecule Dataset Log S Predictions from HF Theory

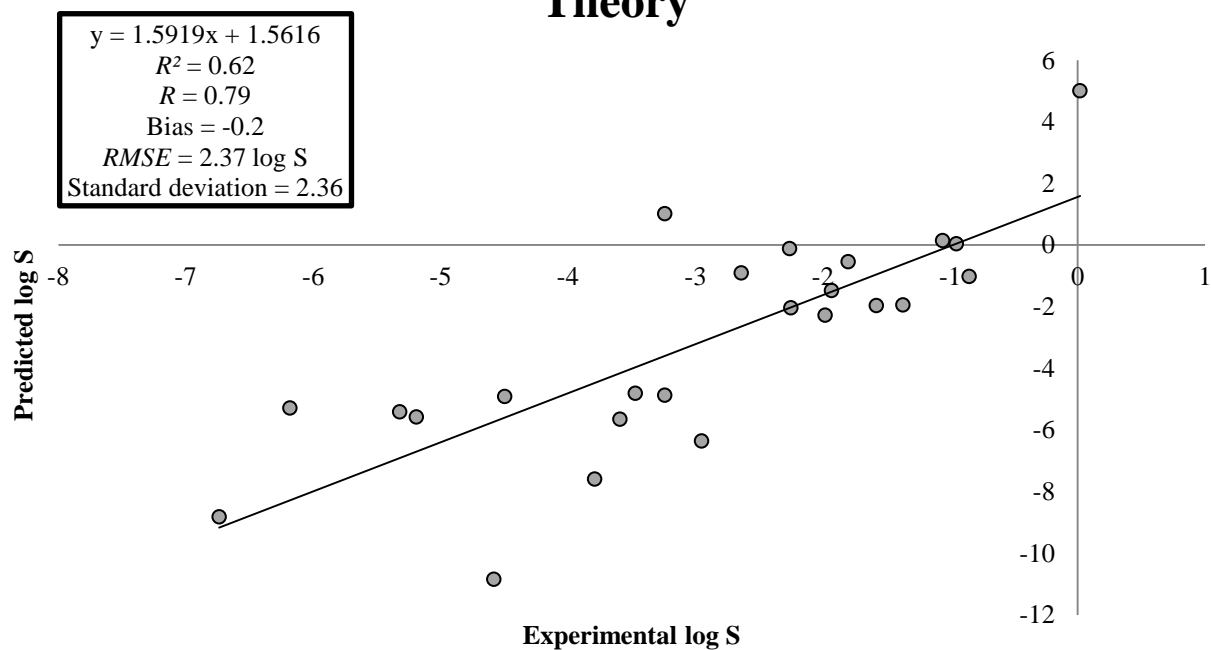


Chart S1: 25 molecule dataset predictions SMD(HF).

Crystal structure	Chemical name	SMILES
ALOPUR	Allopurinol	<chem>c1c2c([nH]n1)ncnc2O</chem>
AMBNAC04	4-Aminobenzoic acid	<chem>O=C(O)c1ccc(N)cc1</chem>
AMXBPM10	Trimethoprim	<chem>COc1cc(cc(c1OC)OC)Cc2cnc(nc2N)N</chem>
BENZAC02	Benzoic acid	<chem>c1ccc(cc1)C(=O)O</chem>
BZAMID02	Benzamide	<chem>c1ccc(cc1)C(=O)N</chem>
COCAIN10	Cocaine	<chem>CN1[C@H]2CC[C@@H]1[C@H]([C@H](C2)OC(=O)c3ccccc3)C(=O)OC</chem>
COYRUD11	Naproxen	<chem>C[C@@H](c1ccc2cc(ccc2c1)OC)C(=O)O</chem>
DHANQU06	1,8-Dihydroxyanthraquinone	<chem>c1cc2c(c(c1)O)C(=O)c3c(cccc3O)C2=O</chem>
EPHPMO	Primidone	<chem>O=C1NCNC(=O)C1(c2ccccc2)CC</chem>
ESTRON14	Estrone	<chem>O=C4[C@]3(CC[C@@H]2c1ccc(O)cc1CC[C@H]2[C@@H]3CC4)C</chem>
HXACAN04	Paracetamol	<chem>CC(=O)Nc1ccc(cc1)O</chem>
IBPRAC01	Ibuprofen	<chem>CC(C)Cc1ccc(cc1)C(C)C(=O)O</chem>
IVUQOF	Fluconazole	<chem>c1cc(c(cc1F)F)C(Cn2cncn2)(Cn3cncn3)O</chem>
JODTUR01	Isoproturon	<chem>O=C(Nc1ccc(cc1)C(C)C)N(C)C</chem>
LABJON01	Nitrofurantoin	<chem>O=[N+](=[O-])c2oc(/C=N/N1C(=O)NC(=O)C1)cc2</chem>
NAPHOL01	1-Naphthol	<chem>Oc2ccccc1ccccc12</chem>
NDNHCL01	Clozapine	<chem>CN1CCN(CC1)C2=Nc3cc(ccc3Nc4c2cccc4)Cl</chem>
NICOAC02	Nicotinic acid	<chem>c1cc(cnc1)C(=O)O</chem>
NIFLUM10	Niflumic acid	<chem>FC(F)(F)c1cc(ccc1)Nc2ncccc2C(=O)O</chem>
PINDOL	Pindolol	<chem>CC(C)NCC(O)COc1cccc2[nH]ccc12</chem>
PTERID11	Pteridine	<chem>n1c2c(ncc1)ncnc2</chem>
PYRENE07	Pyrene	<chem>c1cc2ccc3cccc4c3c2c(c1)cc4</chem>
SALIAC	Salicylic acid	<chem>c1ccc(c(c1)C(=O)O)O</chem>
SIKLIH01	Diclofenac	<chem>c1ccc(c(c1)CC(=O)O)Nc2c(cccc2Cl)Cl</chem>
XYANAC	Mefenamic acid	<chem>O=C(O)c2c(Nc1cccc(c1)C)cccc2</chem>

Table S2: Names, CSD refcodes and SMILES for the 25 molecules in dataset DLS-25.¹³ The full SMILES dataset can be found in the zip file of solubility datasets and scripts that forms part of the Supporting Information.

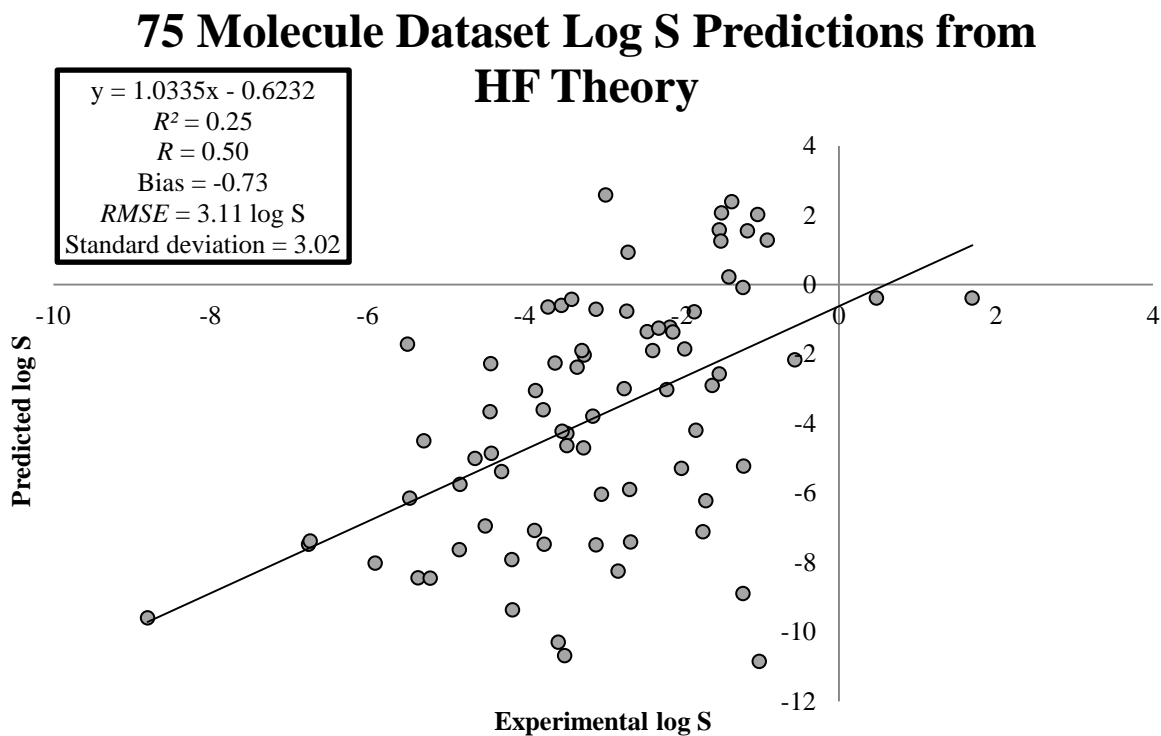


Chart S2: 75 molecule dataset predictions SMD(HF).

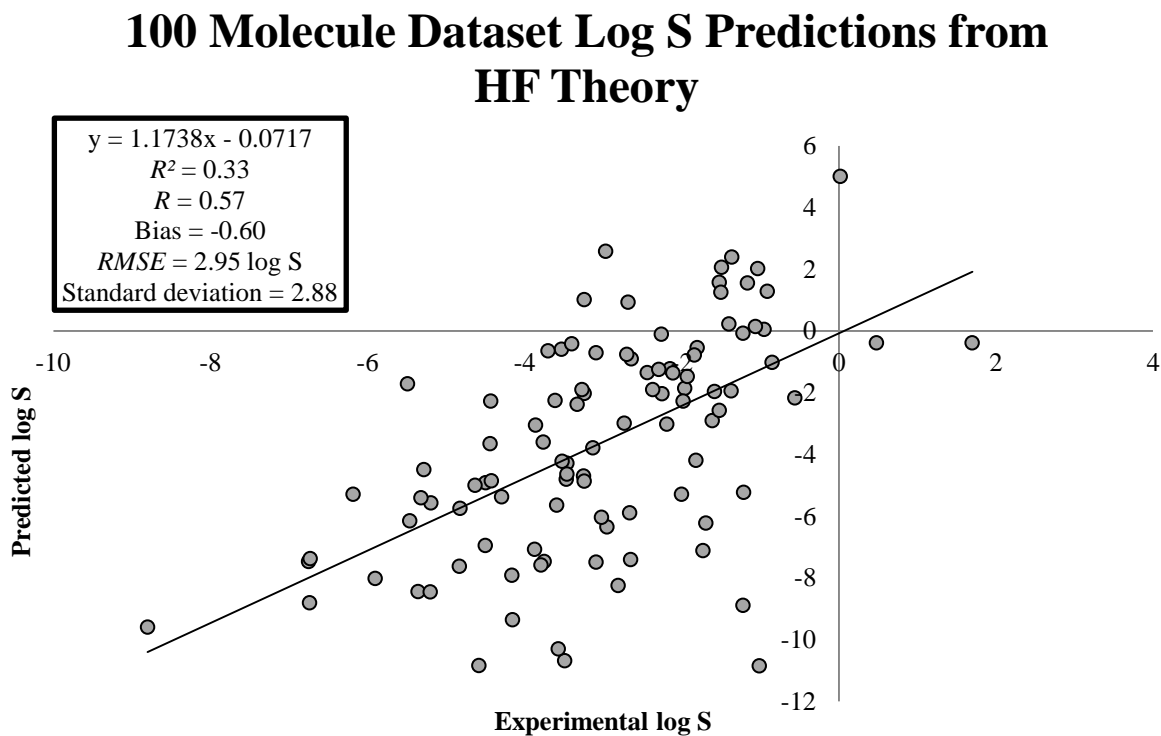


Chart S3: 100 molecule dataset predictions SMD(HF).

25 Molecule Dataset Log S Predictions from DFT M06-2X

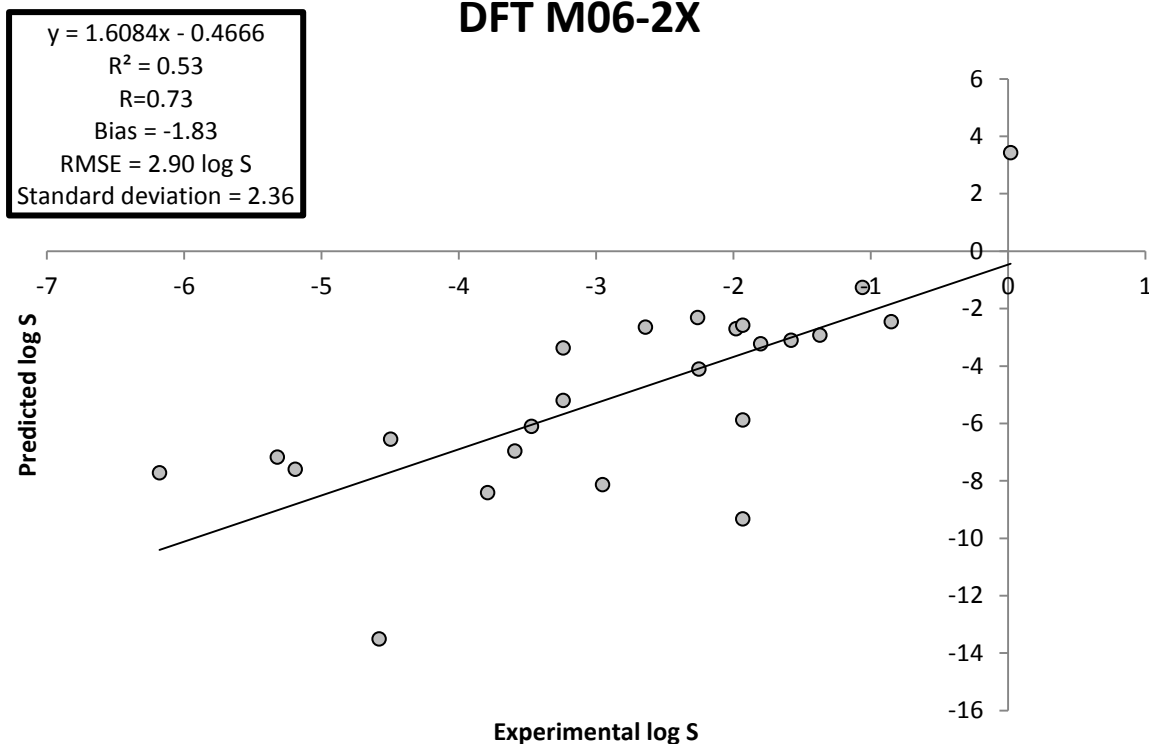


Chart S4: 25 molecule dataset predictions DFT SMD(M06-2X).

75 Molecule Dataset Log S Predictions from DFT M06-2X

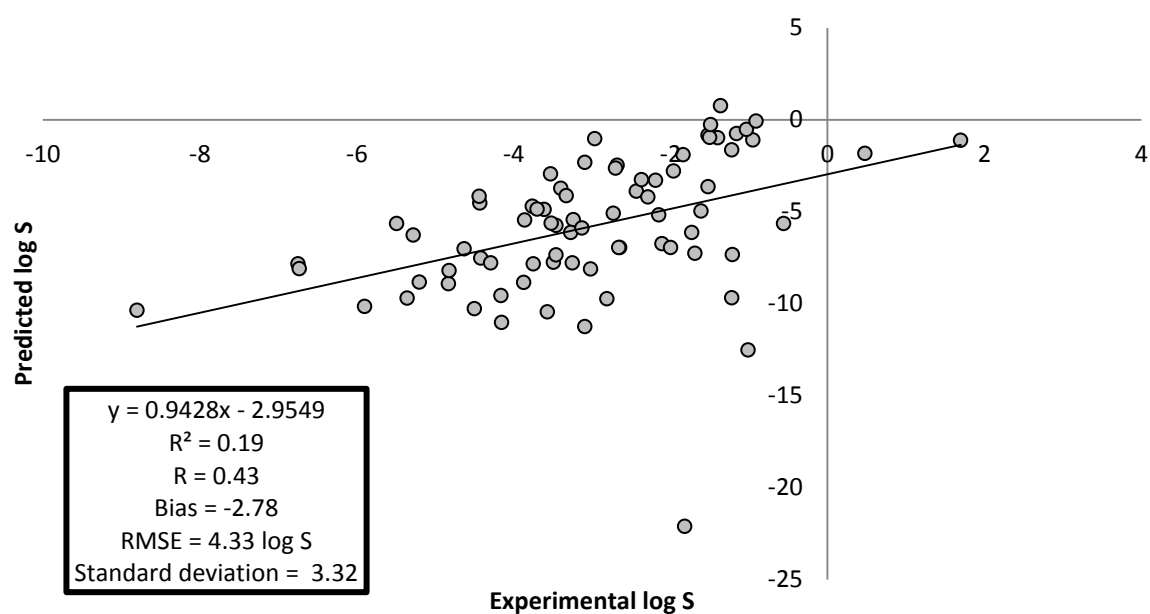


Chart S5: 75 molecule dataset predictions DFT SMD(M06-2X).

100 Molecule Dataset Log S Predictions from DFT M06-2X

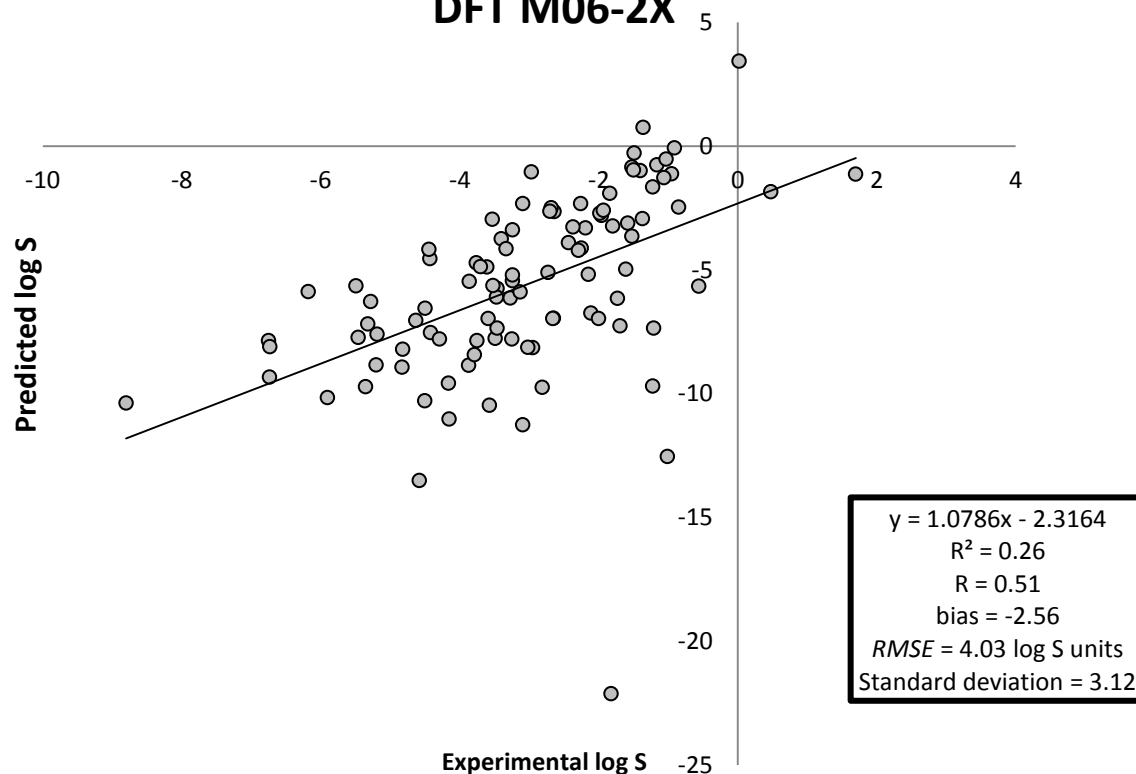


Chart S6: 100 molecule dataset prediction DFT SMD(M06-2X).

Supplementary Results

R^2 results

Informatics Descriptors	Scaled by mean / stdev \pm stdev	Scaled by PCA \pm stdev	Raw data \pm stdev
SVR	0.51 ± 0.02	0.46 ± 0.03	0.46 ± 0.06
RF	0.53 ± 0.02	0.48 ± 0.02	0.53 ± 0.02
PLS	0.52 ± 0.05	0.53 ± 0.01	0.42 ± 0.06

Table S3. Cheminformatics descriptors: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation.

HF Energies learned	Scaled by mean / stdev \pm stdev	Scaled by PCA \pm stdev	Raw data \pm stdev
SVR	0.46 ± 0.02	0.46 ± 0.02	0.46 ± 0.02
RF	0.47 ± 0.03	0.5 ± 0.02	0.47 ± 0.03
PLS	0.36 ± 0.01	0.36 ± 0.02	0.29 ± 0.03

Table S4. Hartree-Fock energy terms: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation obtained when Hartree-Fock energy terms are used as features in machine learning.

HF and Descriptors	Scaled by mean / stdev ± stdev	Scaled by PCA ± stdev	Raw data ± stdev
SVR	0.54 ± 0.03	0.47 ± 0.03	0.44 ± 0.04
RF	0.56 ± 0.02	0.52 ± 0.01	0.56 ± 0.02
PLS	0.57 ± 0.04	0.54 ± 0.03	0.35 ± 0.05

Table S6. Hartree-Fock energy terms and Cheminformatics descriptors: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation.

M062X Energies learned	Scaled by mean / stdev ± stdev	Scaled by PCA ± stdev	Raw data ± stdev
SVR	0.45 ± 0.02	0.46 ± 0.02	0.45 ± 0.02
RF	0.47 ± 0.02	0.4 ± 0.03	0.47 ± 0.02
PLS	0.35 ± 0.02	0.35 ± 0.02	0.25 ± 0.04

Table S5. M06-2X: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation obtained when M06-2X energy terms are used as features in machine learning.

M062X and Descriptors	Scaled by mean / stdev ± stdev	Scaled by PCA ± stdev	Raw data ± stdev
SVR	0.53 ± 0.02	0.46 ± 0.02	0.43 ± 0.05
RF	0.57 ± 0.02	0.54 ± 0.01	0.57 ± 0.02
PLS	0.59 ± 0.02	0.56 ± 0.02	0.35 ± 0.07

Table S7. M06-2X and Cheminformatics descriptors: average $R^2 \pm$ Standard Deviation for the predicted and experimental log S values over ten repetitions of the 10-fold cross-validation.

Solubility Challenge	Scaled by mean / stdev ± stdev	Scaled by PCA ± stdev	Raw data ± stdev
SVR	0.45 ± 0.02	0.31 ± 0.02	0.39 ± 0.04
RF	0.56 ± 0.01	0.36 ± 0.02	0.56 ± 0.01
PLS	0.55 ± 0.02	0.53 ± 0.02	0.33 ± 0.03

Table S9. Solubility Challenge dataset: R^2 for the calculated against experimental log S values for ten repetitions of 10-fold cross-validation using cheminformatics descriptors.

Solubility Challenge	Scaled by mean/stdev	Scaled by PCA	Raw data
SVR	0.41	0.39	0.41
RF	0.50	0.50	0.57
PLS	0.55	0.55	0.58

Table S8. Solubility Challenge dataset: R^2 for the calculated against experimental log S values for the original Solubility Challenge training:test split using cheminformatics descriptors.

RMSE results

Informatics Descriptors	Scaled by mean / stdev \pm stdev	Scaled by PCA \pm stdev	Raw data \pm stdev
SVR	1.19 \pm 0.03	1.25 \pm 0.03	1.25 \pm 0.06
RF	1.17 \pm 0.03	1.24 \pm 0.02	1.17 \pm 0.03
PLS	1.22 \pm 0.1	1.19 \pm 0.02	1.39 \pm 0.1

Table S10. Cheminformatics descriptors: average over ten repetitions of the 10-fold cross-validation of *RMSE* \pm Standard Deviation for the predicted and experimental log S values.

HF Energies learned	Scaled by mean / stdev \pm stdev	Scaled by PCA \pm stdev	Raw data \pm stdev
SVR	1.25 \pm 0.02	1.26 \pm 0.02	1.25 \pm 0.02
RF	1.24 \pm 0.03	1.21 \pm 0.02	1.24 \pm 0.03
PLS	1.37 \pm 0.02	1.36 \pm 0.02	1.45 \pm 0.03

Table S11. Hartree-Fock energy terms: average over ten repetitions of the 10-fold cross-validation of *RMSE* \pm Standard Deviation for the predicted and experimental log S values obtained when HF energy terms are used as features in a machine learning model.

HF and Descriptors	Scaled by mean / stdev \pm stdev	Scaled by PCA \pm stdev	Raw data \pm stdev
SVR	1.16 \pm 0.03	1.25 \pm 0.03	1.28 \pm 0.05
RF	1.14 \pm 0.02	1.19 \pm 0.01	1.14 \pm 0.02
PLS	1.15 \pm 0.06	1.18 \pm 0.04	1.47 \pm 0.08

Table S13. Hartree-Fock energy terms and Cheminformatics descriptors: average *RMSE* \pm Standard Deviation over ten repetitions of the 10-fold cross-validation for the predicted and experimental log S values.

M062X and Descriptors	Scaled by mean / stdev \pm stdev	Scaled by PCA \pm stdev	Raw data \pm stdev
SVR	1.17 \pm 0.02	1.25 \pm 0.02	1.28 \pm 0.05
RF	1.13 \pm 0.02	1.17 \pm 0.01	1.13 \pm 0.02
PLS	1.11 \pm 0.04	1.14 \pm 0.03	1.47 \pm 0.12

Table S12. M06-2X energy terms and Cheminformatics descriptors: average over ten repetitions of the 10-fold cross-validation of *RMSE* \pm Standard Deviation for the predicted and experimental log S values.

M062X Energies learned	Scaled by mean / stdev \pm stdev	Scaled by PCA \pm stdev	Raw data \pm stdev
SVR	1.26 \pm 0.03	1.25 \pm 0.03	1.26 \pm 0.03
RF	1.24 \pm 0.02	1.32 \pm 0.03	1.24 \pm 0.02
PLS	1.37 \pm 0.02	1.38 \pm 0.04	1.51 \pm 0.06

Table S14. M06-2X energy terms: average over ten repetitions of the 10-fold cross-validation of *RMSE* \pm Standard Deviation for the predicted and experimental log S values obtained when M06-2X energy terms are used as features in a machine learning model.

Solubility Challenge	Scaled by mean / stdev \pm stdev	Scaled by PCA \pm stdev	Raw data \pm stdev
SVR	1.03 \pm 0.02	1.15 \pm 0.01	1.08 \pm 0.04
RF	0.93 \pm 0.01	1.12 \pm 0.01	0.93 \pm 0.01
PLS	0.93 \pm 0.02	0.95 \pm 0.02	1.17 \pm 0.04

Table S15. Solubility Challenge dataset: *RMSE* for the calculated against experimental log S values for ten repetitions of 10-fold cross-validation using cheminformatics descriptors.

Solubility Challenge	Scaled by mean/stdev	Scaled by PCA	Raw data
SVR	1.068	1.083	1.079
RF	1.032	1.021	0.927
PLS	0.913	0.913	0.887

Table S16. Solubility Challenge dataset: *RMSE* for the calculated against experimental log S values for the original Solubility Challenge training:test split using cheminformatics descriptors.

Statistical Significance Test

Scaled by the mean and standard deviation						Chemoinformatics descriptors			
Partial Least Square						SVR	RF		PLS
	mx	dd	hf	mx			SVR		
mx	x						x		
hf	0.14	x					0.13	x	
dd	0.19	0.06	x				0.18	0.23	x
hf	0.00	0.02	0.04	x					
mx	0.00	0.02	0.09	0.20	x				
Support Vector Regression						HF + Chemoinformatics Descriptors			
	mx	dd	hf	mx		SVR	RF		PLS
mx	x						x		
hf	0.29	x					0.12	x	
dd	0.36	0.07	x				0.06	0.22	x
hf	0.03	0.04	0.13	x					
mx	0.05	0.06	0.09	0.37	x				
Random Forest Regression						MX06-2X + Chemoinformatics Descriptors			
	mx	dd	hf	mx		SVR	RF		PLS
mx	x						x		
hf	0.26	x					0.03	x	
dd	0.02	0.11	x				0.16	0.28	x
hf	0.00	0.01	0.01	x					
mx	0.02	0.02	0.07	0.25	x				
						HF			
	mx	dd	hf	mx		SVR	RF		PLS
mx	x						x		
hf	0.25	x					0.25	x	
dd	0.02	0.11	x				0.03	0.01	x
hf	0.00	0.01	0.01	x					
mx	0.02	0.02	0.07	0.25	x				
						MX06-2X			
	mx	dd	hf	mx		SVR	RF		PLS
mx	x						x		
hf	0.25	x					0.20	x	
dd	0.02	0.11	x				0.03	0.01	x
hf	0.00	0.01	0.01	x					
mx	0.02	0.02	0.07	0.25	x				
mx = M06-2X + CHEMOINFORMATICS DESCRIPTORS						SVR = SUPPORT VECTOR REGRESSION			
hf = HF + CHEMOINFORMATIC DESCRIPTORS						RF = RANDOM FOREST			
dd = CHEMOINFORMATICS DESCRIPTORS						PLS = PARTIAL LEAST SQUARE			
hf = HF									
mx = MX06-2X									

BOX S1: P-value (statistical significance at $P = 0.05$) of the performance of the *RMSE* scores for the different regression models for the scaled dataset by using mean/stdev.

Principal components						Chemoinformatics descriptors			
Partial Least Square									
	mx	hf	dd	hfd	mx	SVR	RF	PLS	
mx	x					x			
hfd	0.18	x				0.41	x		
dd	0.11	0.15	x			0.20	0.23	x	
hf	0.00	0.01	0.01	x					
mx	0.00	0.02	0.01	0.11	x				
Support Vector Regression						HF + Chemoinformatics Descriptors			
	mx	hf	dd	hfd	mx	SVR	RF	PLS	
mx	x					x			
hfd	0.31	x				0.15	x		
dd	0.23	0.08	x			0.13	0.25	x	
hf	0.09	0.19	0.12	x					
mx	0.05	0.16	0.19	0.23	x				
Random Forest Regression						MX06-2X + Chemoinformatics Descriptors			
	mx	hf	dd	hfd	mx	SVR	RF	PLS	
mx	x					x			
hfd	0.10	x				0.02	x		
dd	0.01	0.08	x			0.06	0.08	x	
hf	0.19	0.20	0.38	x					
mx	0.01	0.01	0.10	0.07	x				
mx = M06-2X + CHEMOINFORMATICS DESCRIPTORS hfd = HF + CHEMOINFORMATIC DESCRIPTORS dd = CHEMOINFORMATICS DESCRIPTORS hf = HF mx = MX06-2X						HF			
						SVR	RF	PLS	
						x			
						0.11	x		
						0.01	0.00	x	
						MX60-2X			
						SVR	RF	PLS	
						x			
						0.15	x		
						0.04	0.26	x	
						SVR = SUPPORT VECTOR REGRESSION RF = RANDOM FOREST PLS = PARTIAL LEAST SQUARE			

BOX S2: P-value (statistical significance at $P = 0.05$) of the performance of the *RMSE* scores for the different regression models for the scaled dataset by Principal Components.

Raw data set						Chemoinformatics descriptors			
Partial Least Square									
	mx	hf	dd	hf	mx	SVR	RF	PLS	
mx	x					x			
hf	0.03	x				0.10	x		
dd	0.19	0.17	x			0.10	0.05	x	
hf	0.11	0.13	0.21	x					
mx	0.17	0.24	0.23	0.27	x				
Support Vector Regression						HF + Chemoinformatics Descriptors			
	mx	hf	dd	hf	mx	SVR	RF	PLS	
mx	x					x			
hf	0.28	x				0.06	x		
dd	0.24	0.29	x			0.07	0.01	x	
hf	0.06	0.22	0.11	x					
mx	0.09	0.14	0.20	0.37	x				
Random Forest Regression						MX06-2X + Chemoinformatics Descriptors			
	mx	hf	dd	hf	mx	SVR	RF	PLS	
mx	x					x			
hf	0.23	x				0.07	x		
dd	0.02	0.16	x			0.17	0.02	x	
hf	0.01	0.01	0.01	x					
mx	0.02	0.02	0.07	0.25	x				
mx = M06-2X + CHEMOINFORMATICS DESCRIPTORS						HF			
hf = HF + CHEMOINFORMATIC DESCRIPTORS									
dd = CHEMOINFORMATICS DESCRIPTORS						SVR	RF	PLS	
hf = HF						x			
mx = MX06-2X						0.25	x		
						0.01	0.00	x	
						MX60-2X			
	mx	hf	dd	hf	mx	SVR	RF	PLS	
mx	x					x			
hf	0.20	x				0.20	x		
dd	0.01	0.01	x			0.01	0.01	x	
hf				x					
mx				0.25	x				
						SVR = SUPPORT VECTOR REGRESSION			
						RF = RANDOM FOREST			
						PLS = PARTIAL LEAST SQUARE			

BOX S3: P-value (statistical significance at $P = 0.05$) of the performance of the *RMSE* scores for the different regression models for the row dataset.

Variable Importance

Top 10 variables Ranking of variable importance in Random Forest Scaled by mean/stdev (stdev)				
Descriptor only	Descriptor and HF	Descriptor and M06-2X	HF	M06-2X
XLogP	XLogP	XLogP	dG.solv	dG.solv
WTPT.3	WTPT.3	DFT_logS	HF_logS	dG.solution
VCH.7	DFT.logS	dG.solution	dG.solution	DFT_logS
ATSc2	dG.solution	WTPT.3	dGsub	Srot
SP.6	VCH.7	VCH.7	Ulatt	Strans
ATSc1	dG.solv	dG.solv	Scrys	Soln energy
SP.5	ATSc1	ATSc1	Srot	Ulatt
SP.7	SP.6	ATSc2	Strans	Scrys
ATSm4	ATSc2	WTPT.2	Soln energy	Gas energy
ATSm1	WTPT.2	SP.6	Gas energy	dGsub

Top 10 variables Ranking of variable importance in Random Forest Raw data				
Descriptor only	Descriptor and HF	Descriptor and M06-2X	HF	M06-2X
XLogP	XLogP	XLogP	dG.solv	dG.solv
WTPT.3	WTPT.3	dG.solution	HF_logS	dG.solution
VCH.7	DFT.logS	DFT.logS	dG.solution	DFT_logS
ATSc2	dG.solution	WTPT.3	dGsub	Srot
ATSc1	VCH.7	dG.solv	Ulatt	Strans
SP.6	dG.solv	VCH.7	Scrys	Soln energy
SP.5	ATSc1	ATSc1	Srot	Ulatt
ATSm5	ATSc2	ATSc2	Strans	Scrys
ATSm4	SP.6	WTPT.2	Soln energy	Gas energy
SP.7	SP.5	SP.6	Gas energy	dGsub

Table S17: Top 10 results of variable importance for different descriptors and dataset.

References

1. Zhao, Y.; Truhlar, D., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120* (1), 215-241.

2. Wheeler, S. E.; Houk, K. N., Integration Grid Errors for Meta-GGA-Predicted Reaction Energies: Origin of Grid Errors for the M06 Suite of Functionals. *J. Chem. Theory Comput.* **2010**, 6 (2), 395-404.
3. Max Born; Huang, K., *Dynamical Theory of Crystal Lattices* Oxford University press: New York, 1954.
4. Krishnan, A.; Williams, L. J.; McIntosh, A. R.; Abdi, H., Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage* **2011**, 56 (2), 455-475.
5. Debasish Basak; Srimanta Pal; Patranabis, D. C., Support Vector Regression. *Neural Information Processing - Letters and Reviews* **2007**, 203 - 224.
6. Nath, N.; Mitchell, J. B. O., Is EC class predictable from reaction mechanism? *BMC Bioinformatics* **2012**, 13 (1), 60.
7. Menke, J.; Martinez, T. R. In Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons, *Neural Netw.*, 2004. Proceedings. 2004 IEEE International Joint Conference on, 25-29 July 2004; 2004; pp 1331-1335 vol.2.
8. Kuhn, M. Variable Importance Using The caret Package 2010, p. 1-7.
9. Kuhn, M.; Contributions from Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T. CARET: Classification and Regression Training 2013. <http://CRAN.R-project.org/package=caret>.
10. (a) CDK Descriptor Summary (2011-05-28). <http://pele.farmbio.uu.se/nightly-1.2.x/dnames.html>; (b) CDK Small Molecule Descriptors. http://tyerslab.bio.ed.ac.uk/wikiworld/sga/index.php/CDK_Small_Molecule_Descriptors#AutoCorrelationDescriptorPolarizability.
11. Llinàs, A.; Glen, R. C.; Goodman, J. M., Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008**, 48 (7), 1289-1303.
12. (a) The Goodman group. <http://www-jmg.ch.cam.ac.uk/data/solubility/> (accessed 8 February 2013); (b) Hopfinger, A. J.; Esposito, E. X.; Llinàs, A.; Glen, R. C.; Goodman, J. M., Findings of the Challenge To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, 49 (1), 1-5.
13. Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; Fedorov, M. V., First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Theory Comput.* **2012**, (8), 3322-3337.